Department of Digital Business
Journal of Artificial Intelligence and Digital Business (RIGGS)
Homepage: https://journal.ilmudata.co.id/index.php/RIGGS
Vol. 4 No. 2 (2025) pp: 821-832
P-ISSN: 2963-9298, e-ISSN: 2963-914X

Evaluasi Kinerja SVM dan *Logistic Regression* pada Data *Multiclass* dalam Analisis Sentimen Film *Dirty Vote* dengan Metode Pelabelan *Lexicon Based*

Andini Sintawati[1], Farida Amalya[2], Ahmad Hidayat[3], Nur Muhammad Supyan[4]
[1,2,3,4]Jurusan Informatika, Fakultas Teknologi Industri, Universitas Gunadarma
Email: [1]aries@staff.gunadarma.ac.id, [2]faridaamalya@gmail.com, [3]ahmad_hidayat@staff.gunadarma.ac.id, [4]nmsupyan@gmail.com

Abstrak
Pada 11 Februari 2024, saluran Youtube Dirty Vote dan PSHK Indonesia merilis film berjudul Dirty Vote yang menuai perdebatan masyarakat. Dalam rangka mengetahui sentimen masyarakat, dilakukan analisis sentimen. Analisis sentimen merupakan metode untuk mengkategorikan sentimen dengan melibatkan Natural Language Processing (NLP) dan algoritma machine learning, seperti Support Vector Machine (SVM) dan Logistic Regression. Penelitian sebelumnya telah membandingkan kedua algoritma tersebut dalam melakukan analisis sentimen pada dua atau tiga kategori. Namun, pada penelitian tiga kategori SVM hanya dilatih dan diuji dengan kernel RBF. Oleh karena itu, dilakukan penelitian untuk membandingkan nilai akurasi model SVM dan Logistic Regression dalam mengklasifikasikan sentimen film Dirty Vote. SVM dilatih dan diuji menggunakan tiga kernel, yaitu Polynomial, RBF, dan Sigmoid. Penelitian ini menggunakan tahapan-tahapan NLP dengan menggunakan data sebanyak 3.500 yang berasal dari proses scraping film Dirty Vote. Data digolongkan menjadi tiga kategori, yaitu sentimen negatif sebesar 36,31%, sentimen positif sebesar 31,95%, dan sentimen netral sebesar 31,74%. Data tersebut dibagi menjadi data latih dan data uji dengan rasio 90:10, 80:20, 70:30. Dari penelitian ini, diperoleh hasil rata-rata akurasi untuk ketiga kernel SVM dan Logistic Regression. Hasil penelitian menunjukan nilai akurasi tertinggi diperoleh oleh algoritma SVM dengan kernel RBF sebesar 77%, diikuti oleh Logistic Regression sebesar 76%, kernel Sigmoid sebesar 75%, dan kernel Polynomial sebesar 65%.

Kata Kunci: Multiclass, SVM, Logistic Regresion, Kernel, Dirty Vote

1. Latar Belakang

Dampak dari berkembangnya media massa memudahkan masyarakat dalam mendapatkan informasi yang dibutuhkan. Selain memudahkan dalam mengakses informasi, media massa juga dapat digunakan untuk memberikan pendidikan, hiburan, dan pengaruh kepada masyarakat. Salah satu dari media massa adalah film (Christa, Deisy, dan Grace, 2021).

Pada tanggal 11 Februari 2024, saluran Youtube *Dirty Vote* dan PSHK Indonesia merilis sebuah film berjudul *Dirty Vote* (Maryam et. al, 2024). Film tersebut memicu perdebatan di kalangan masyarakat dikarenakan memaparkan dugaan-dugaan kecurangan yang terjadi dalam pemilihan presiden. Akibatnya, film Dirty Vote tersebut menimbulkan pertentangan dan dukungan dari berbagai pihak (Diah, Bucky, dan Dudi, 2024).

Perdebatan mengenai film *Dirty Vote* memenuhi kolom komentar dari kedua saluran Youtube. Dengan jumlah komentar yang sangat banyak, maka akan sangat sulit untuk mengetahui sentimen masyarakat terhadap film tersebut. Oleh sebab itu, analisis sentimen diperlukan untuk mengetahui sentimen masyarakat terhadap film tersebut.

Analisis sentimen merupakan metode yang biasa digunakan dalam menilai dan mengkategorikan sebuah sentimen seperti positif, negatif, atau netral (Ismia, Agung, dan Gatot, 2023). Proses analisis sentimen melibatkan penggunaan metode pada pemrosesan bahasa alami dan algoritma yang digunakan dalam pembelajaran mesin seperti *Naive Bayes*, *Support Vector Machine* (SVM), dan *K-Nearest Neighbor* (KNN) (Sisferi, Amsal, dan Siti, 2020). Selain ketiga algoritma tersebut, terdapat algoritma lain seperti *Decision Tree* dan *Logistic Regression*.

Penelitian terdahulu yang berkaitan adalah penelitian yang berjudul Analisis Sentimen Pengguna Aplikasi *Marketplace* Tokopedia Pada Situs Google Play Menggunakan *Support Vector Machine* (SVM), *Naive Bayes*, dan *Logistic Regression*. Penelitian tersebut menggunakan data dari hasil *scraping* web Google Play dan dibagi menjadi dua

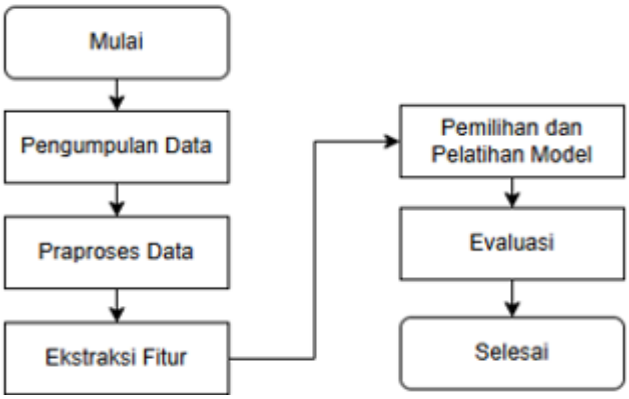
kategori. Hasil dari penelitian tersebut menghasilkan akurasi pada *Support Vector Machine* (SVM) sebesar 98%, *Naive Bayes* sebesar 83%, dan *Logistic Regression* sebesar 86% (Meishita dan Iqbal, 2022).

Penelitian lain yang berkaitan adalah penelitian berjudul Analisis Sentimen Persepsi Publik Terhadap UPN "Veteran" Jawa Timur Menggunakan Metode SVM, *Naïve Bayes*, dan *Multinomial Logistic Regression*. Data penelitian diperoleh melalui proses *scraping* pada Twitter. Data dikategorikan menjadi tiga kategori, yaitu positif, netral, dan negatif menggunakan *TextBlob*. Data dilatih dan diuji menggunakan algoritma *Naïve Bayes*, SVM dengan kernel RBF, dan *Multinomial Logistic Regression* dengan rasio pembagian data 80:20. Dari penelitian tersebut, algoritma SVM dengan kernel RBF memperoleh nilai akurasi sebesar 66%, *Naïve Bayes* sebesar 72%, dan *Multinomial Logistic Regression sebesar* 75% (Rahmatul, Sahat, Prismahardi, 2023).

Berdasarkan uraian di atas, maka diketahui bahwa perbandingan SVM dan *Logistic Regression* dalam analisis sentimen untuk dua atau tiga kategori telah dilakukan. Namun, pada penelitian yang membandingkan tiga kategori, algroritma SVM hanya menggunakan kernel RBF. Oleh karena itu, dilakukan penelitian yang bertujuan untuk membandingkan algoritma SVM dan *Logistic Regression* pada tiga label kategori, yaitu positif, negatif, dan netral. Pada SVM data dilatih dan diuji dengan menggunakan tiga kernel, yaitu *Polynomial*, RBF, dan *Sigmoid*. Penelitian ini juga dilakukan untuk mengetahui sentimen publik terhadap film *Dirty Vote*.

2. Metode Penelitian

Penelitian ini melalui beberapa proses tahapan pemrosesan bahasa alami, yaitu pengumpulan data, praproses data, ekstraksi fitur, pemilihan dan pelatihan model, dan evaluasi. Tahapan dari pemrosesan bahasa alami ditampilkan melaui gambar 1.



Gambar 1. Tahapan Penelitian

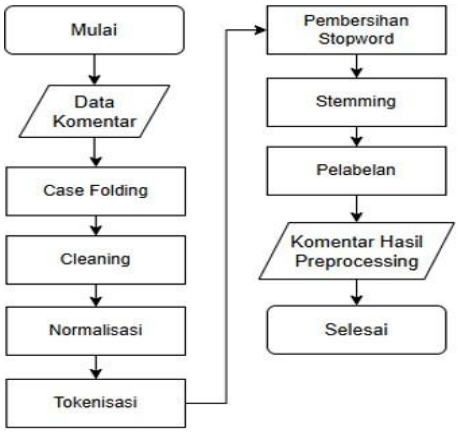
2.1 Pengumpulan Data

Pada tahapan ini dilakukan pengumpulan data dari komentar film *Dirty Vote* yang didapatkan dari saluran Youtube *Dirty Vote* sebagai sumber data pada penelitian ini. Alasan pemilihan tersebut dikarenakan saluran Youtube tersebut merupakan salah satu saluran awal yang merilis film tersebut.

2.2 Praproses Data

Pada tahapan praproses data dilakukan pembersihan dan penyiapan data agar menjadi terstruktur untuk diproses pada tahapan selanjutnya. Tahapan praproses data secara umum diawali dengan tahapan *case folding*, pembersihan, tokenisasi, penghapusan *stopword*, dan *stemming*. Namun, pada penelitian ini ditambahkan dua buah tahapan, yaitu tahapan normalisasi dan pelabelan. Pada tahapan normalisasi dilakukan sebelum melakukan tahapan tokenisasi, sedangkan tahapan pelabelan dilakukan setelah tahapan *stemming* dilakukan. Alur dari tahapan praproses data dapat dilihat pada gambar 2.

DOI: https://doi.org/10.31004/riggs.v4i2.576

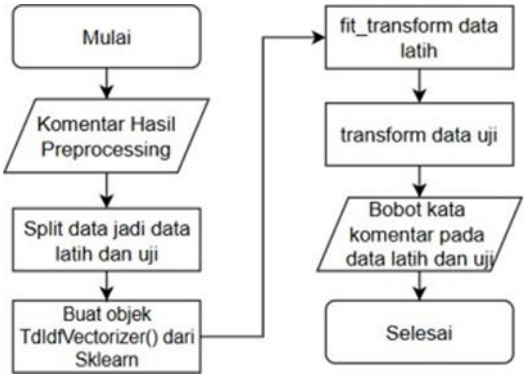


Gambar 2. Tahapan Praproses Data

Tahapan praproses data terdiri dari beberapa langkah penting untuk menyiapkan data sebelum digunakan dalam pelatihan model. Proses dimulai dengan case folding, yaitu mengubah seluruh kata menjadi huruf kecil agar seragam. Selanjutnya dilakukan cleaning, yaitu membersihkan data dari tag dan atribut HTML, link, tanda baca, kata satu huruf, dan kata berulang yang bersebelahan. Tahap berikutnya adalah normalisasi, yaitu mengubah kata singkatan, kata gaul, dan kesalahan ketik menjadi bentuk baku. Setelah itu dilakukan pembersihan stopword untuk menghapus kata-kata yang tidak memberikan informasi signifikan bagi model. Proses kemudian dilanjutkan dengan stemming, yaitu mengubah kata berimbuhan menjadi bentuk dasarnya. Terakhir adalah tahapan pelabelan, di mana data mentah diberi label sentimen menggunakan kamus SentiStrength Indonesia dengan metode *Lexicon Based*, yakni kamus hasil translasi dari bahasa Inggris oleh Devid Haryalesmana Wahid dan Azhari SN dalam penelitiannya tahun 2016.

2.3 Ekstraksi Fitur

Pada tahapan ekstraksi fitur dilakukan suatu proses untuk mengolah sebuah data mentah menjadi format yang dapat diolah pada saat pelatihan model. Format tersebut berbentuk representasi numerik yang didapatkan dari hasil pembobotan frekuensi pada kata. Pembobotan frekuensi pada kata dilakukan dengan menggunakan metode TF-IDF dengan menggunakan fungsi *TfIdfVectorizer()* dari pustaka scikit- learn. Pembobotan frekuensi pada kata ditampilkan dalam bentuk sebuah diagram alir yang dapat dilihat pada gambar 3.



Gambar 3. Tahapan Ekstraksi Fitur

2.4 Pemilihan dan Pelatihan Model

Pada tahapan pemilihan dan pelatihan model terjadi proses pelatihan dan pengujian pada algoritma SVM dan Logistic Regression. Pada SVM data dilatih dengan tiga buah kernel, yaitu RBF, Polynomial, dan Sigmoid. Pada algoristma SVM dalam menangani data yang memiliki lebih dari dua kategori dibutuhkan metode

DOI: https://doi.org/10.31004/riggs.v4i2.576

pendekatan khusus, yaitu OVR dan OVO. Pada penelitian ini metode pendekatan yang digunakan adalah OVO, dimana metode tersebut membuat model-model kecil sebanyak jumlah label dan dilakukan perhitungan algoritma SVM pada tiap model. Kumpulan model kecil tersebut digabungkan menjadi satu buah model utama SVM. Berikut disajikan sebuah tabel yang menggambarkan cara kerja model-model kecil tersebut pada label target. Tabel tersebut dapat dilihat pada tabel 1.

Tabel 1. Cara Kerja Metode One-Against-All

Model Biner	$y_i = 1$	$y_i = -1$
Model Biner 1 (Positif vs Netral dan Negatif)	0	1 dan 2
Model Biner 2 (Netral vs Positif dan Negatif)	1	0 dan 2
Model Biner 3 (Negatif vs Positif dan Netral)	2	0 dan 1

Sementara itu, pada algoritma Logistic Regression untuk menangani data multiclass, maka digunakan pendekatan multinomial atau yang lebih dikenal dengan nama Multinomial Logistic Regression. Pada pendekatan multinomial maka dibuat persamaan logit sebanyak K+1 kategori. Sehingga jika terdapat tiga buah kategori, maka membutuhkan dua buah persamaan logit. Persamaan logit dapat dilihat pada persamaan 1. berikut.

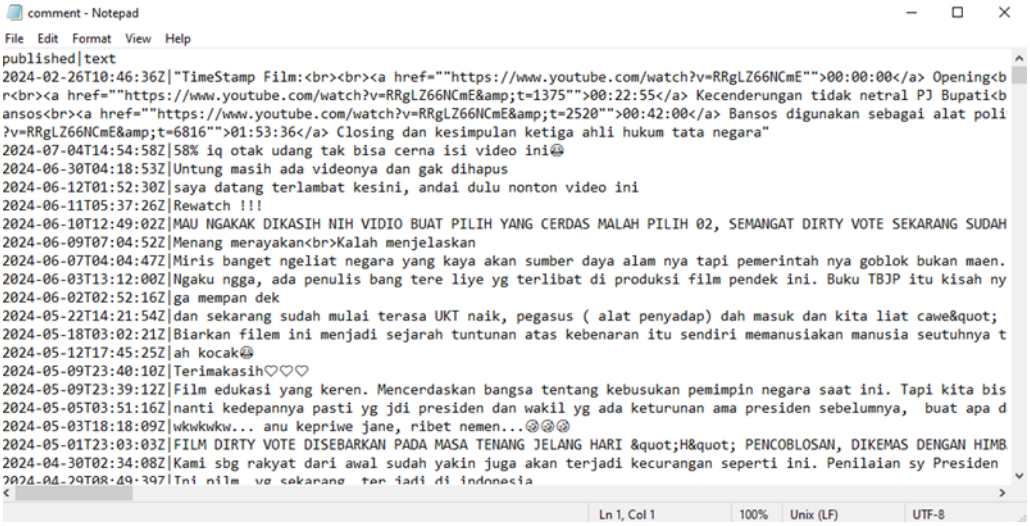
$$f(k, i) = \beta_{0,k} + \beta_{1,k}x_{1i} + \beta_{2,k}x_{2i} + \cdots + \beta_{m,k}x_{mi} = x^T\beta_k \quad (1)$$
$$i$$

2.5 Evaluasi

Proses tahapan evaluasi dilakukan dengan menggunakan sebuah fungsi bernama *classification_report()* dari scikit-learn. Fungsi tersebut digunakan untuk menampilkan laporan klasifikasi dari tiap model yang dilatih. Dari hasil laporan klasifikasi pada fungsi tersebut, maka dapat dilakukan perbandingan antar algoritma.

3. Hasil dan Diskusi

Hasil dari pengumpulan data diperoleh 3.500 data dalam rentang tanggal 16 Februari 2024 hingga 16 Juli 2024. Data tersebut didapatkan dari hasil *scraping* komentar film *Dirty Vote* pada saluran Youtube *Dirty Vote* dengan bantuan Youtube APIv3. Hasil dari pengumpulan data dapat dilihat pada gambar 4.



Gambar 4. Hasil Pengumpulan Data

3.1 Praproses Data

Tahapan praproses data melalui beberapa tahapan di dalamnya, seperti *case folding*, pembersihan data, normalisasi, tokenisasi, pembersihan stopword, *stemming*, dan pelabelan. Pada hasil yang diperoleh dari tahapan *case folding* menghasilkan seluruh kata pada komentar yang telah dikumpulkan pada tahapan pengumpulan data menjadi huruf kecil semua. Hal tersebut dikarenakan kata yang sama dengan huruf kecil dan huruf besar dianggap berbeda oleh komputer. Hasil dari proses yang terjadi pada tahapan *case folding* dapat dilihat pada gambar 5.

DOI: https://doi.org/10.31004/riggs.v4i2.576

	published	text
0	2024-02-26T10:46:36Z	timestamp film: <a href="https://www.yo...
1	2024-07-04T14:54:58Z	58% iq otak udang tak bisa cerna isi video ini😂
2	2024-06-30T04:18:53Z	untung masih ada videonya dan gak dihapus
3	2024-06-12T01:52:30Z	saya datang terlambat kesini, andai dulu nonto...
4	2024-06-11T05:37:26Z	rewatch !!!
5	2024-06-10T12:49:02Z	mau ngakak dikasih nih vidio buat pilih yang c...
6	2024-06-09T07:04:52Z	menang merayakan kalah menjelaskan
7	2024-06-07T04:04:47Z	miris banget ngeliat negara yang kaya akan sum...
8	2024-06-03T13:12:00Z	ngaku ngga, ada penulis bang tere liye yg terl...
9	2024-06-02T02:52:16Z	ga mempan dek

Gambar 5. Hasil *Case Folding*

Pada tahapan *cleaning*, data dibersihkan dari tanda baca, tag dan atribut HTML, link, kata yang hanya satu huruf, dan kata sama yang bersebelahan. Tahapan *cleaning* juga membuang beberapa baris data yang duplikat dan kosong. Sehingga data yang awalnya berjumlah 3.500 data menjadi 3.333 data. Hasil dari tahapan cleaning seperti pada gambar 6.

	published	text
1	2024-07-04T14:54:58Z	iq otak udang tak bisa cerna isi video ini
2	2024-06-30T04:18:53Z	untung masih ada videonya dan gak dihapus
3	2024-06-12T01:52:30Z	saya datang terlambat kesini andai dulu nonton...
4	2024-06-11T05:37:26Z	rewatch
5	2024-06-10T12:49:02Z	mau ngakak dikasih nih vidio buat pilih yang c...
6	2024-06-09T07:04:52Z	menang merayakan kalah menjelaskan
7	2024-06-07T04:04:47Z	miris banget ngeliat negara yang kaya akan sum...
8	2024-06-03T13:12:00Z	ngaku ngga ada penulis bang tere liye yg terli...
9	2024-06-02T02:52:16Z	ga mempan dek
10	2024-05-22T14:21:54Z	dan sekarang sudah mulai terasa ukt naik pegas...

Gambar 6. Hasil Tahapan *Cleaning*

Pada tahapan normaliasi dilakukan proses mengubah kata gaul, singkatan, kata yang salah ketik menjadi kata yang sebenarnya. Pada penelitian ini, proses normalisasi membutuhkan kamus yang berisi 770 daftar kata yang perlu diperbaiki. Daftar kata tersebut disimpan ke dalam file normalization.txt. Hasil dari tahapan normalisasi dapat dilihat pada gambar 7.

	published	text
1	2024-07-04T14:54:58Z	iq otak udang tidak bisa cerna isi video ini
2	2024-06-30T04:18:53Z	untung masih ada videonya dan tidak dihapus
3	2024-06-12T01:52:30Z	saya datang terlambat ke sini andai dulu nonto...
4	2024-06-11T05:37:26Z	rewatch
5	2024-06-10T12:49:02Z	mau ngakak dikasih nih video buat pilih yang p...
6	2024-06-09T07:04:52Z	menang merayakan kalah menjelaskan
7	2024-06-07T04:04:47Z	miris banget ngeliat negara yang seperti akan ...
8	2024-06-03T13:12:00Z	ngaku tidak ada penulis bapak tere liye yang t...
9	2024-06-02T02:52:16Z	tidak mempan dek
10	2024-05-22T14:21:54Z	dan sekarang sudah mulai terasa ukt naik pegas...

Gambar 7. Hasil Tahapan Normalisasi

Tahapan tokenisasi menghasilkan data komentar yang telah dipecah-pecah menjadi bagian-bagian yang lebih kecil seperti kata. Hasil dari tahapan tokenisasi dapat dilihat pada gambar 8.

DOI: https://doi.org/10.31004/riggs.v4i2.576

	published	text
0	2024-07-04T14:54:58Z	[iq, otak, udang, tidak, bisa, cerna, isi, vid...
1	2024-06-30T04:18:53Z	[untung, masih, ada, videonya, dan, tidak, dih...
2	2024-06-12T01:52:30Z	[saya, datang, terlambat, ke, sini, andai, dul...
3	2024-06-11T05:37:26Z	[rewatch]
4	2024-06-10T12:49:02Z	[mau, ngakak, dikasih, nih, video, buat, pilih...
5	2024-06-09T07:04:52Z	[menang, merayakan, kalah, menjelaskan]
6	2024-06-07T04:04:47Z	[miris, banget, ngeliat, negara, yang, seperti...
7	2024-06-03T13:12:00Z	[ngaku, tidak, ada, penulis, bapak, tere, liye...
8	2024-06-02T02:52:16Z	[tidak, mempan, dek]
9	2024-05-22T14:21:54Z	[dan, sekarang, sudah, mulai, terasa, ukt, nai...

Gambar 8. Hasil Tahapan Tokenisasi

Pada tahapan pembersihan *stopword* menghasilkan sebuah data komentar yang telah bersih dari kata-kata *stopword*. Hal tersebut dikarenakan pada tahapan ini terjadi proses penghapusan kata-kata *stopword*, seperti kata hubung dan kata-kata yang dianggap tidak memberikan banyak informasi pada saat pelatihan model. Hasil tahapan pembersihan *stopword* yang dapat dilihat pada gambar 9.

	published	text
0	2024-07-04T14:54:58Z	[iq, otak, udang, tidak, cerna, isi, video]
1	2024-06-30T04:18:53Z	[untung, videonya, tidak, dihapus]
2	2024-06-12T01:52:30Z	[terlambat, andai, nonton, video]
3	2024-06-11T05:37:26Z	[rewatch]
4	2024-06-10T12:49:02Z	[ngakak, dikasih, video, pilih, pintar, pilih,...
5	2024-06-09T07:04:52Z	[menang, merayakan, kalah]
6	2024-06-07T04:04:47Z	[miris, banget, ngeliat, negara, sumber, daya,...
7	2024-06-03T13:12:00Z	[ngaku, tidak, penulis, tere, liye, terlibat, ...
8	2024-06-02T02:52:16Z	[tidak, mempan]
9	2024-05-22T14:21:54Z	[ukt, pegasus, alat, penyadap, masuk, lihat, c...

Gambar 9. Hasil Tahapan Pembersihan *Stopword*

Proses tahapan pembersihan *stopword* mengurangi jumlah data dari 3.333 data menjadi 3.308 data. Hal tersebut dikarenakan proses pada *stopword* memungkinkan beberapa baris data menjadi kosong, sehingga baris data tersebut perlu dibersihkan. Setelah tahapan pembersihan *stopword* dilakukan, dilakukan tahapan *stemming* yang bertujuan untuk menghasilkan bentuk kata dasar untuk setiap kata pada data komentar. Hasil dari tahapan *stemming* dapat dilihat pada gambar 10.

	published	text
0	2024-07-04T14:54:58Z	[iq, otak, udang, tidak, cerna, isi, video]
1	2024-06-30T04:18:53Z	[untung, video, tidak, hapus]
2	2024-06-12T01:52:30Z	[lambat, andai, nonton, video]
3	2024-06-11T05:37:26Z	[rewatch]
4	2024-06-10T12:49:02Z	[ngakak, kasih, video, pilih, pintar, pilih, s...
5	2024-06-09T07:04:52Z	[menang, raya, kalah]
6	2024-06-07T04:04:47Z	[miris, banget, ngeliat, negara, sumber, daya,...
7	2024-06-03T13:12:00Z	[ngaku, tidak, tulis, tere, liye, libat, produ...
8	2024-06-02T02:52:16Z	[tidak, mempan]
9	2024-05-22T14:21:54Z	[ukt, pegasus, alat, sadap, masuk, lihat, cawe...

Gambar 10. Hasil Tahapan *Stemming*

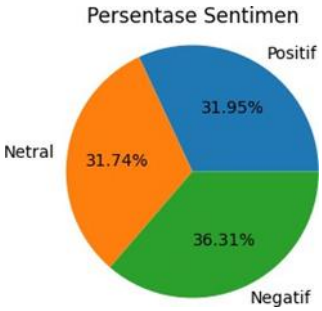
Tahapan pelabelan dilakukan untuk melabeli data komentar yang masih berbentuk data mentah. Pelabelan dilakukan dengan kamus SentiStrangth Indonesia menggunakan metode *Lexicon Based*. Data positif dilabeli dengan angka 0, data netral dilabeli dengan 1, sedangkan data negatif dilabeli dengan 2. Hasil dari tahapan pelabelan dapat dilihat pada gambar 11.

DOI: https://doi.org/10.31004/riggs.v4i2.576

	published	text	score	sentiment
0	2024-07-04T14:54:58Z	[iq, otak, udang, tidak, cerna, isi, video]	-4	2
1	2024-06-30T04:18:53Z	[untung, video, tidak, hapus]	0	1
2	2024-06-12T01:52:30Z	[lambat, andai, nonton, video]	-3	2
3	2024-06-11T05:37:26Z	[rewatch]	0	1
4	2024-06-10T12:49:02Z	[ngakak, kasih, video, pilih, pintar, pilih, s...	5	0
5	2024-06-09T07:04:52Z	[menang, raya, kalah]	0	1
6	2024-06-07T04:04:47Z	[miris, banget, ngeliat, negara, sumber, daya,...	-6	2
7	2024-06-03T13:12:00Z	[ngaku, tidak, tulis, tere, liye, libat, produ...	-10	2
8	2024-06-02T02:52:16Z	[tidak, mempan]	0	1
9	2024-05-22T14:21:54Z	[ukt, pegasus, alat, sadap, masuk, lihat, cawe...	4	0

Gambar 11. Hasil Tahapan Pelabelan

Hasil pelabelan pada *dataset* diperoleh hasil jika data pada didominasi oleh sentimen negatif sebesar 36,31%, sentimen positif sebesar 31,95%, dan sentimen netral sebesar 31,74%. Persentase dari tiap kategori label dapat dilihat pada gambar 12.



Gambar 12. Persentase Tiap Kategori Label

3.2 Ekstraksi Fitur

Pada tahapan ekstraksi fitur terjadi proses pengubahan data teks menjadi representasi numerik yang didapatkan dari hasil pembobotan pada kata. Sebelum dilakukan pembobotan, data komentar dibagi menjadi data latih dan data uji dengan rasio 70:30, 80:20. dan 90:10 seperti pada tabel 2.

Tabel 2. Rasio Pembagian Data Latih dan Uji

Rasio	Data Latih	Data Uji
90 : 10	2.977	331
80 : 20	2.647	662
70 : 30	2.316	993

Setelah data dibagi menjadi data latih dan data uji, maka dilakukan pembobotan frekuensi pada data dengan menggunakan metode TF-IDF. Hasil dari pembobotan dengan menggunakan metode TF-IDF seperti pada gambar 13.

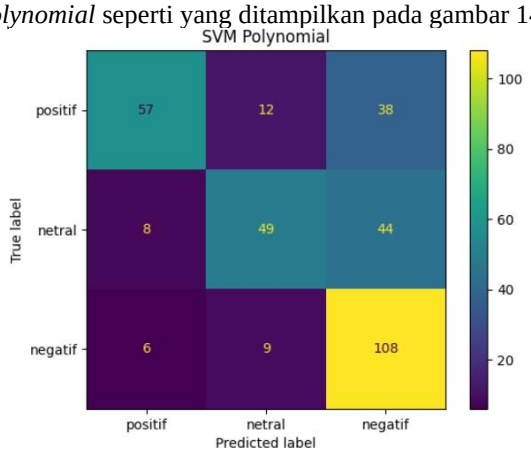
```
(0, 5100)     0.26615103285980113
(0, 1856)     0.31955614301353896
(0, 843)      0.4782183426393777
(0, 4820)     0.14307399704529916
(0, 5005)     0.4782183426393777
(0, 3470)     0.36566290275521385
(0, 1843)     0.46420082465951956
(1, 3000)     0.2019150622218928
(1, 3644)     0.27940686665136066
(1, 4512)     0.4544016824017929
(1, 824)      0.4544016824017929
(1, 3463)     0.37855950741712296
(1, 1899)     0.320784381181492
```

Gambar 13. Hasil Ekstraksi Fitur

DOI: https://doi.org/10.31004/riggs.v4i2.576

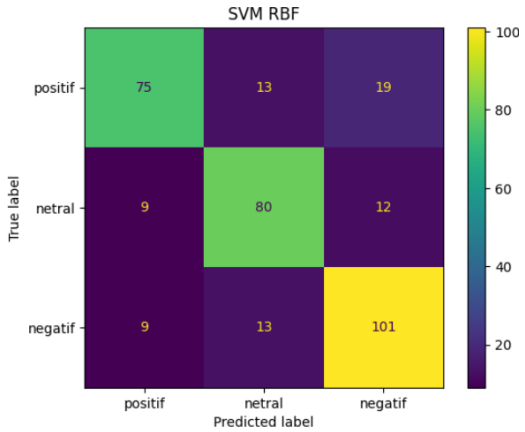
3.3 Hasil Klasifikasi Algoritma

Data yang telah diubah menjadi bentuk numerik dengan menggunakan metode TF-IDF digunakan dalam melakukan klasifikasi dengan menggunakan algoritma SVM dan *Logistic Regression*. Hasil dari klasifikasi divisualisasikan dengan menggunakan *confusion matrix*. Pada rasio pembagian data 90:10, hasil klasifikasi SVM pada kernel *Polynomial* seperti yang ditampilkan pada gambar 14.



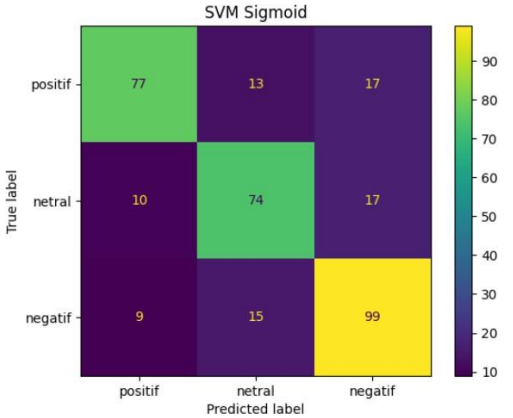
Gambar 14. *Confusion Matrix* Kernel *Polynomial* Rasio 90:10

Demikian pula untuk kernel RBF, hasil klasifikasi SVM dengan rasio 90:10 divisualisasikan dengan menggunakan *confusion matrix*. Hasil *confusion matrix* untuk kernel RBF dengan rasio 90:10 ditampilkan pada gambar 15.



Gambar 15. *Confusion Matrix* Kernel RBF Rasio 90:10

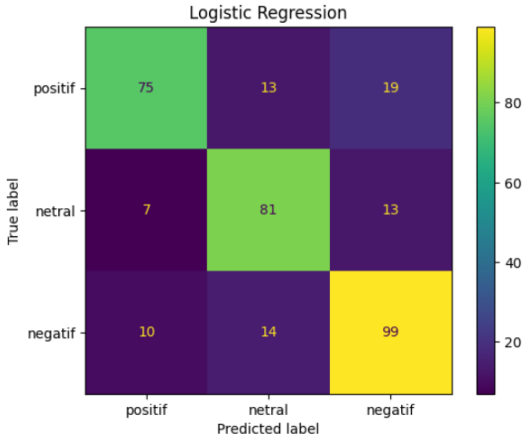
Kemudian, hasil klasifikasi SVM untuk kernel *Sigmoid* dengan rasio 90:10 divisualisasikan dengan menggunakan *confusion matrix*. Berikut *confusion matrix* untuk kernel *Sigmoid* dengan rasio 90:10 yang dapat dilihat pada gambar 16.



Gambar 16. *Confusion Matrix* Kernel *Sigmoid* Rasio 90:10

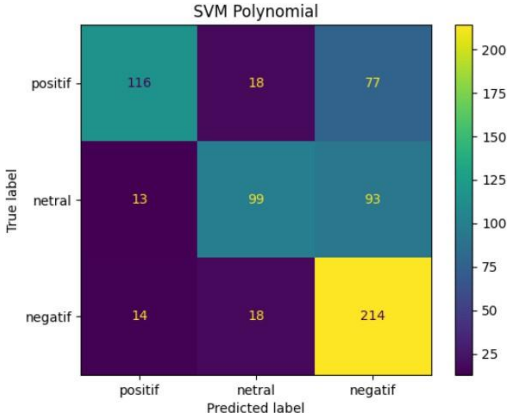
DOI: https://doi.org/10.31004/riggs.v4i2.576

Sementara itu, *confusion matrix* untuk klasifikasi *Logistic Regression* pada rasio pembagian data 90:10 ditampilkan pada gambar 17.



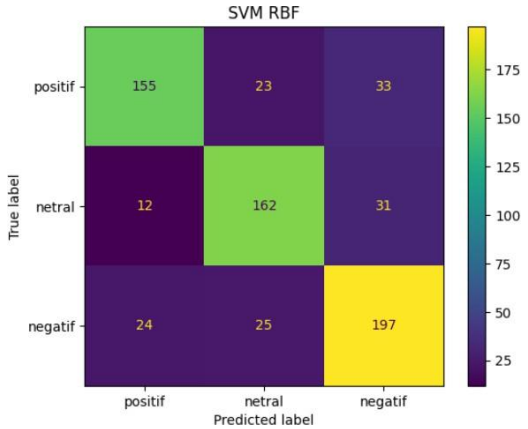
Gambar 17. *Confusion Matrix Logistic Regression* Rasio 90:10

Pada rasio pembagian data 80:20, hasil klasifikasi SVM pada kernel *Polynomial* seperti yang ditampilkan pada gambar 17.



Gambar 18. Confusion Matrix Kernel Polynomial Rasio 80:20

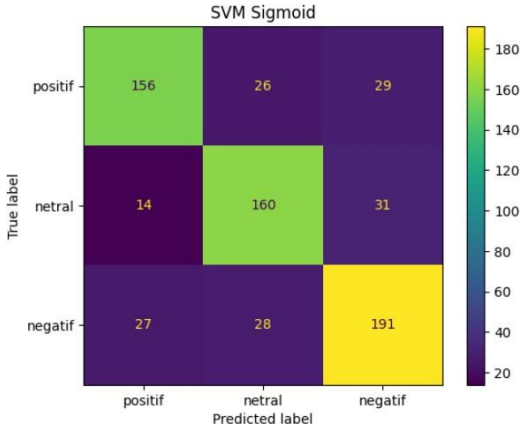
Hal yang sama juga diterapkan pada kernel RBF. Hasil dari klasifikasi SVM dengan rasio 80:20 pada kernel RBF ditampilkan dengan menggunakan *confusion matrix* yang dapat dilihat pada gambar 19.



Gambar 19. Confusion Matrix Kernel RBF Rasio 80:20

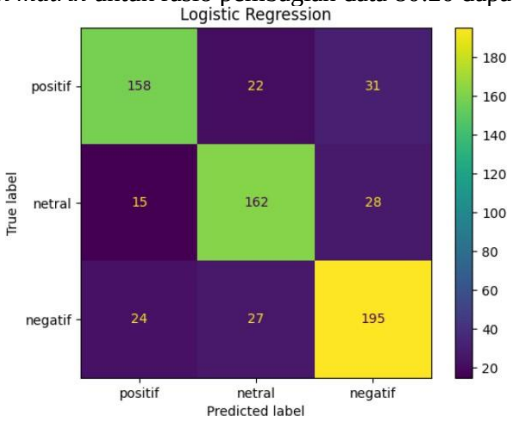
Kemudian, *confusion matrix* untuk kernel *Sigmoid* pada algoritma SVM dengan menggunakan rasio pembagian data 80:20 ditampilkan pada gambar 20.

DOI: https://doi.org/10.31004/riggs.v4i2.576



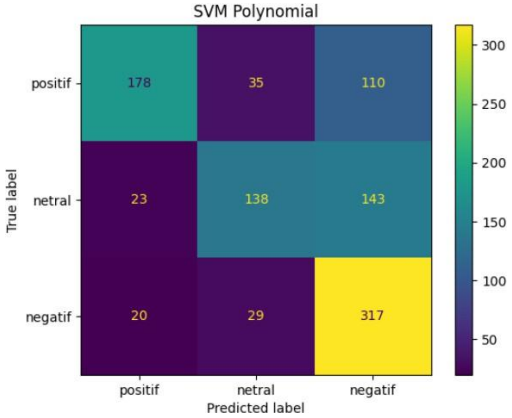
Gambar 20. *Confusion Matrix* Kernel *Sigmoid* Rasio 80:20

Hal yang sama juga dilakukan pada kedua rasio pembagian data yang lainnya pada algoritma *Logistic Regression*. Visualisasi *confusion matrix* untuk rasio pembagian data 80:20 dapat dilihat pada gambar 21.



Gambar 21. *Confusion Matrix Logistic Regression* Rasio 80:20

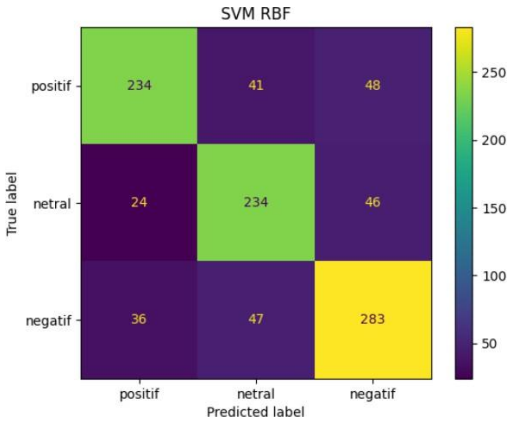
Pada klasifikasi SVM dengan menggunakan rasio pembagian data 70:30 divisualisasikan menggunakan *confusion matrix*. Hasil *confusion matrix* untuk kernel *Polynomial* dengan rasio pembagian data 70:30 yang ditampilkan pada gambar 22.



Gambar 22. *Confusion Matrix* Kernel *Polynomial* Rasio 70:30

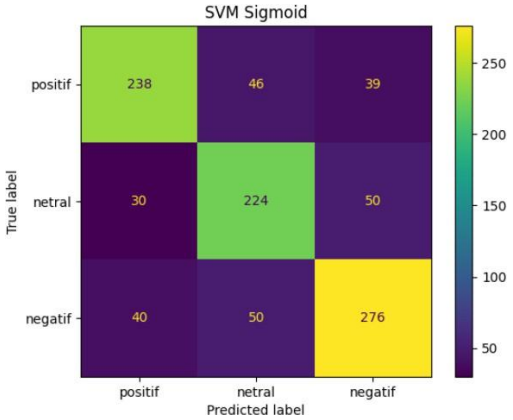
Hal yang sama diterapkan untuk kedua kernel yang lainnya. Pada visualisasi *confusion matrix* untuk kernel RBF dapat dilihat pada gambar 23.

DOI: https://doi.org/10.31004/riggs.v4i2.576



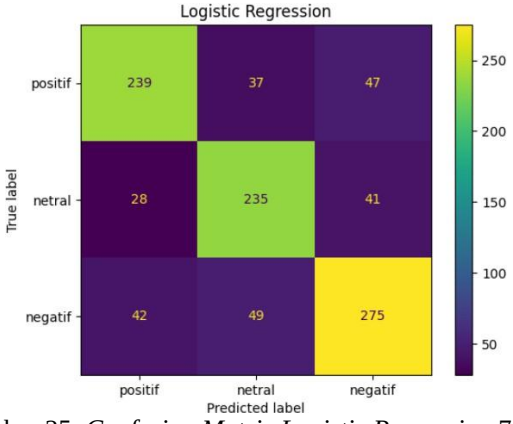
Gambar 23. *Confusion Matrix* Kernel RBF Rasio 70:30

Kemudian, visualisasi *confusion matrix* untuk kernel *Sigmoid* pada rasio pembagian data 70:30 ditampilkan pada gambar 24.



Gambar 24. *Confusion Matrix* Kernel *Sigmoid* Rasio 70:30

Sementara itu, visualisasi dengan *confusion matrix* pada rasio pembagian data 70:30 ditampilkan pada gambar 25.



Gambar 25. *Confusion Matrix Logistic Regression* 70:30

3.4 Evaluasi

Pada tahapan evaluasi, hasil dari *confusion matrix* digunakan untuk mengetahui nilai presisi, akurasi, *recall*, dan *f1-score*. Hasil evaluasi tersebut menggambarkan seberapa baik kedua model klasifikasi tersebut dalam mengklasifikasikan data komentar. Dari nilai akurasi yang didapatkan pada hasil evaluasi, dapat dilakukan perbandingan akurasi dari kedua model algoritma, yaitu *Support Vector Machine* (SVM) dan *Logistic Regression*. Pada SVM digunakan tiga buah kernel, yaitu RBF, *Polynomial*, dan *Sigmoid*. Dengan melakukan perbandingan akurasi dari kedua algoritma tersebut, maka dapat diketahui algoritma yang lebih baik untuk

DOI: https://doi.org/10.31004/riggs.v4i2.576

analisis sentimen data *multiclass* pada film *Dirty Vote*. Perbandingan akurasi dari kedua algoritma disajikan dalam bentuk tabel yang dapat dilihat pada tabel 3.

Tabel 3. Perbandingan Akurasi SVM dan Logistic Regression

	Support Vector Machine (SVM)			*Logistic*
	Polynomial	RBF	*Sigmoid*	*Regression*
90 : 10	0,65	0,77	0,76	0,77
80 : 20	0,65	0,78	0,77	0,78
70 : 30	0,64	0,76	0,74	0,75
Rata-rata	0,64	0,77	0,75	0,76

4. Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa penerapan algoritma Support Vector Machine (SVM) dengan kernel Radial Basis Function (RBF) menunjukkan tingkat akurasi tertinggi sebesar 77% dalam analisis sentimen film *Dirty Vote* dibandingkan dengan Logistic Regression, kernel Sigmoid, dan kernel Polynomial. Proses praproses data yang meliputi case folding, cleaning, normalisasi, pembersihan stopword, stemming, hingga pelabelan menggunakan metode *Lexicon Based* terbukti efektif dalam mempersiapkan data untuk klasifikasi sentimen multiclass. Hasil ini menunjukkan bahwa pemilihan algoritma dan kernel yang tepat berperan penting dalam meningkatkan performa klasifikasi sentimen pada data yang bersumber dari media sosial atau platform daring. Oleh karena itu, penelitian selanjutnya menggunakan teknik optimasi, seperti SMOTE dan ADASYN untuk meningkatkan akurasi. Hal tersebut dikarenakan berdasarkan penelitian yang telah dilakukan, klasifikasi sentimen film *Dirty Vote* dengan tiga kategori label menggunakan algoritma SVM dan *Logistic Regression* hanya menghasilkan akurasi dalam rentang 65% hingga 78%.

Referensi

Christha, A., Warouw, D. M. D., & Waleleng, G. J. (2021). Pesan moral pada film Cek Toko Sebelah (analisis semiotika John Fiske). *Acta Diurna Komunikasi*, *3*(4), 1–7.

Devid, H. W., & Azhari, S. N. (2016). Peringkasan sentimen ekstraktif di Twitter menggunakan hybrid TF-IDF dan cosine similarity. *Indonesian Journal of Computing and Cybernetics Systems*, *10*(2), 207–218. https://doi.org/10.22146/ijccs.16625

Diah, F. S., Guna, B. W. K., & Yudhakusuma, D. (2024). Analisis sentimen film *Dirty Vote* menggunakan BERT (Bidirectional Encoder Representations from Transformers). *Jurnal Teknologi Informasi dan Komunikasi*, *8*(2), 393–404. https://doi.org/10.35870/jtik.v8i2.1580

Ismia, I., Triayudi, A., & Soepriyono, G. (2023). Analisa sentimen pengguna transportasi Jakarta terhadap Transjakarta menggunakan metode Naïve Bayes dan K-Nearest Neighbor. *Journal of Information System Research (JOSH)*, *4*(2), 543–550. https://doi.org/10.47065/josh.v4i2.293

Meishita, I. P., & Kharisudin, I. (2022). Analisis sentimen pengguna marketplace Tokopedia pada situs Google Play menggunakan metode Support Vector Machine (SVM), Naïve Bayes, dan Logistic Regression. *Prosiding Seminar Nasional Matematika (PRISMA)*, *5*, 759–766.

Maryam, S., Sari, A. R., Florid, M. I., Widyastuti, & Runtu, A. R. (2024). Documentary film *Dirty Vote*: Substance and sensation. *International Journal of Society Reviews (INJOSER)*, *2*(4), 956–962.

Sisferi, H., Pardamean, A., & Khasanah, S. N. (2020). Sentimen analisis publik terhadap Joko Widodo terhadap wabah Covid-19 menggunakan metode machine learning. *Jurnal Kajian Ilmiah*, *20*(2), 167–176.

DOI: https://doi.org/10.31004/riggs.v4i2.576