JURNAL ILMIAH FIFO **P-ISSN 2085-4315 / E-ISSN 2502-8332**
DOI: http://dx.doi.org/10.22441/fifo.2022.v14i1.007

Analisis Sentimen Terhadap Pembelajaran Daring Di Indonesia Menggunakan Support Vector Machine (SVM)

Alfiyah Nur Indraini*[1],Iin Ernawati[2]

[1,2]*Jurusan Informatika, Fakultas Ilmu Komputer, Universitas Pembangunan Veteran Jakarta*
Jl. Rs. Fatmawati, Pondok Labu, Jakarta Selatan, DKI Jakarta, 12450, Indonesia
E-mail: [1]alfiyahni@upnvj.ac.id , [2]iinernawati@upnvj.ac.id
*) Korespondensi author

(***received*:** 22-12-2021, ***revised*:** 27-05-2022, ***accepted*:** 28-05-2022)

Abstract

During this pandemic, a new policy was created in the world of education. The policy encourages students to carry out online learning for a long period of time. The new policy raises a lot of public opinion conveyed through social media. Twitter social media is used as a forum for opinions, one of which is about online learning. Therefore, this study will conduct a sentiment analysis on public opinion regarding online learning in Indonesia to provide information or evaluation of public opinion on Twitter social media. Sentiment analysis can be done by classifying public opinion into positive opinion and negative opinion with the Support Vector Machine (SVM) method. In classifying data, data labeling and data cleaning can be carried out first before going through the text preprocessing process, then the data is given a weight for each word with Term Frequency–Inverse Document Frequency (TF-IDF) which will be used as a feature after that the data is divided using a 10-fold cross. validation and classified by the Support Vector Machine (SVM) method. The average results of the evaluation using the confusion matrix are accuracy of 0.72.

***Keyword**: Sentiment Analysis, Classification, Support Vector Machine (SVM).*

Abstrak

Pada masa pandemi ini tercipta kebijakan baru dalam dunia pendidikan. Kebijakan tersebut menganjurkan pelajar untuk melaksanakan pembelajaran dalam jaringan (daring) dengan jangka waktu yang panjang. Kebijakan baru menimbulkan banyaknya opini publik yang disampaikan melalui media sosial. Oleh karena itu, penelitian ini akan melakukan analisis sentimen terhadap opini publik mengenai pembelajaran daring di Indonesia untuk memberikan informasi atau evaluasi terhadap opini publik pada media sosial twitter. Analisis sentimen dapat dilakukan dengan mengklasifikasi opini publik menjadi opini positif dan opini negatif dengan metode Support Vector Machine (SVM). Dalam mengklasifikasikan data dapat dilakukan pelabelan data dan pembersihan data terlebih dahulu sebelum melalui proses text preprocessing, kemudian data diberikan bobot setiap kata dengan Term Frequncy–Invers Document Frequency (TF-IDF) yang akan dijadikan sebagai fitur setelah itu pembagian data menggunakan 10-fold cross validation dan diklasifikasikan dengan metode Support Vector Machine (SVM). Hasil rata-rata evaluasi dengan confusion matrix yaitu accuracy sebesar 0,72.

Kata kunci: Analisis Sentimen, Klasifikasi, Support Vector Machine (SVM).

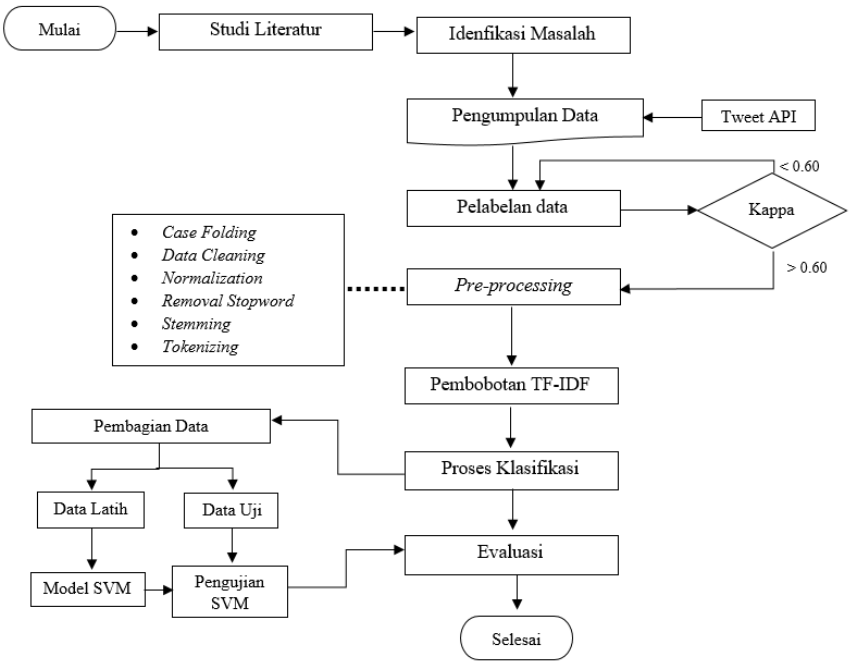
I. Pendahuluan

Saat ini seluruh dunia termasuk Indonesia sedang terdampak sebuah virus yang mematikan yaitu *Covid*-19. Pada masa pandemi ini mengharuskan semua rakyat di Indonesia untuk menjalankan beberapa kebijakan. Kebijakan – kebijakan tersebut memaksa kita untuk menjalankan kebiasaan baru dalam pencegahan rantai penyebaran virus, salah satunya "dirumah aja". Hal ini menimbulkan sebuah kebijakan baru dalam dunia pendidikan yaitu pada Surat Edaran yang dikeluarkan Mendikbud Nomor 4 Tahun 2020, mengenai proses BDR (Belajar Dari Rumah)[1]. Kebijakan tersebut menjelaskan tetang pelaksanan proses Belajar Dari Rumah (BDR) dengan mengubah

pembelajaran tatap muka yang biasa dilaksanakan disekolah maupun dikampus dengan pembelajaran secara *online* atau daring [1]. Dengan dianjurkannya seluruh pelajar di Indonesia untuk tetap belajar dirumah atau disebut dengan PJJ (Pembelajaran Jarak Jauh) memang bukan hal yang baru untuk dilakukan. Namun untuk waktu jangka panjang dan dilakukan secara bersamaan oleh semua tingkat bawah hingga tingkat atas pelajar Indonesia pembelajaran daring merupakan hal yang baru dilakukan. Dengan perubahan sistem pembelajaran jarak jauh ini, menimbulkan beberapa opini publik yang berupa opini positif, opini netral maupun opini negatif dari berbagai kalangan pelajar di indonesia.

Dengan adanya opini publik dapat dilakukan analisis sentimen untuk mendapatkan sebuah informasi. Analisis sentimen merupakan penelitian untuk mengenali opini seseorang terhadap suatu kejadian. Analisis sentimen dilakukan dengan mengelompokan/ klasifikasi polaritas apakah pendapat yang terdapat dalam suatu teks maupun dokumen, kalimat besifat positif, netral maupun negatif [2]. Dalam melakukan klasifikasi pada kumpulan data *tweet* pada media sosial *twitter* harus dilakukan pengolahan teks atau disebut dengan *text mining*. *Text mining* dapat didefinisikan sebagai proses penggalian data berbasis teks yang berkaitan dengan informasi apa yang akan *user* cari serta mencari hubungan atau korelasi yang menarik dengan menggunakan *tools* analisis yang merupakan komponen-komponen dari data mining [3]. Pengolahan teks atau *text mining* dapat dilakukan dengan beberapa tahap dalam *preprocessing* atau pembersihan data teks untuk dapat dilakukan ke tahap klasifikasi berikutnya. Dalam analisis sentimen opini publik dapat diklasifikasikan menjadi opini positif maupun opini negatif agar data dapat menghasilkan sebuah informasi. Algoritma SVM merupakan teknik klasifikasi yang dapat digunakan untuk mengklasifikasikan data. Pada penelitian ini Algoritma klasifikasi yang digunakan adalah metode SVM (*Support vector Machine*). Dengan dilakukan analisis sentimen dapat pada penelitian ini dapat meberikan hasil berupa informasi mengenai proses klasifikasi atau hasil klasifikasi yang dapat dijadikan sebagai evaluasi kefektifan mekanisme pelaksanaan proses Belajar Dari Rumah (BDR) menurut pandangan dimasyarakat pada media sosial *Twitter*.

II. Metodologi Penelitian



Gambar 1. Alur penelitian

Sebelum penelitian, dilakukan identifikasi masalah yang terjadi terkait dengan topik-topik pada penelitian. Kemudian dilakukan studi literatur terkait sebagai sumber tertulis dalam melakukan penelitian. Studi literatur dalam penelitian ini dilakukan dengan mengumpulkan buku-buku, jurnal terkait mengenai analisis sentimen, *text mining*, *pre-processing* dan algoritma klasifikasi *Support Vector Machine* yang dibahas dalam penelitian dengan *tweet* yang mengandung kata "kelas *online*" dan "kelas daring". Opini publik yang digunakan pada analisis sentimen pada penelitian ini menggunakan media sosial *Twitter*. Dalam pengumpulan data *tweet* ini, media sosial

DOI: http://dx.doi.org/10.22441/fifo.2022.v14i1.007

twitter sudah menyediakan API (*Application Programming Interface*) sebagai akses dalam pengambilan data pada *twitter*. Setelah pengumpulan data dilakukan pelabelan data agar data dapat diklasifikasi menjadi dua jenis. yaitu opini publik positif dan opini public negatif. Pelabelan data tweet dilakukan secara manual dengan menggunakan tiga orang sebagai anatator yang akan melabelkan data *tweets* yang akan dianalisis. Pelabelan akan disepakati menggunakan pengukuran tingkat kesepakatan antar anatator dengan *tools fleiss kappa*. *Fleiss kappa* merupakan metode yang digunakan untuk mengukur sebuah kesepakatan dalam melakukan penelitian dengan menggunakan dua atau lebih penilai [4]. Berikut merupakan rumus perhitungan kappa:

$$kappa = \frac{p_\alpha - p_\varepsilon}{1 - p_\varepsilon} \qquad (1)$$

Dengan rumus perhitungan presentase jumlah pengukuran antar rater atau (p_α) sebagai berikut:

$$p_\alpha = \frac{1}{n}\sum_{i=1}^{n} \frac{\sum_{j=1}^{k} x_{ij}^2 - m}{m(m-1)} \qquad (2)$$

dan rumus perhitungan presentase jumlah perubahan antar rater atau (p_ε) sebagai berikut:

$$p_\varepsilon = \sum_{j=1}^{k} q_j^2 \text{ , dimana } q_j \text{ adalah } \quad q_j = \frac{1}{nm}\sum_{i=1}^{n} x_{ij} \qquad (3)$$

Keterangan:

p_α : Presentase jumlah pengukuran antar rater
p_ε : Presentase jumlah perubahan antar rater
n : Jumlah data
m : Jumlah rater
x_{ij} : Jumlah keputusan rater

Berikut merupakan tabel interpretasi kappa:

Tabel 1. Interpretasi Kappa

Indeks Kappa	Agreement
1.00 – 0.81	*Almost perfect agreement*
0.80 – 0.61	*Subtantial agreement*
0.60 – 0.41	*Moderate agreement*
0.40 – 0.21	*Fair agreement*
0.20 – 0.00	*Slight agreement*
<0	*Poor agreement or no agreement*

Pada Tabel 1. untuk indeks kappa sekitar 0,41 hingga 0,60 kesepakatan dapat dikatakan "*Moderate"* atau sedang, sehingga pada penelitian ini 0,60 dijadikan sebagai batas indeks kappa dalam pelabelan data untuk melanjutkan proses penelitian ketahap berikutnya yaitu tahap *Pre-processing*. *Pre-processing* merupakan sebuah tahap awal dalam pengolahan data yang berfungsi untuk membersihkan data sehingga terbentuk sebuah data sesuai kebutuhan untuk dapat diproses lebih lanjut. Data teks yang diperoleh biasanya tidak terstruktur atau data mentah yang memiliki banyak *noise* seperti tanda baca, simbol-simbol, imbuhan , angka dan lain sebagainya [2]. *Text preprocessing* bertujuan untuk mengubah suatu data teks yang tidak terstruktur menjadi sebuah data teks yang terstruktur. Dalam melakukan *text preprocessing* perlu dilakukanya beberapa tahap agar data menjadi sebuah data teks yang terstruktur, Berikut merupakan tahapan *text preprocessing* [5]:

a. *Case Folding*, merupakan tahapan untuk mengubah bentuk kata menjadi sama dengan melakukan perubahan huruf kapital (*uppercase*) menjadi huruf kecil (*lowercase*).
b. *Data Cleaning*, merupakan tahapan yang digunakan untuk menghapus karakter yang telah ditentukan. Karakter yang dihapus seperti *username*, *hastag* dan url dari situs web.
c. *Normalization*, merupakan proses perbaikan kata-kata yang salah eja atau disingkat dalam bentuk tertentu. Pada tahap *normalization* ini bertujuan untuk memperkecil dimensi kata yang memiliki arti sama tetapi memiliki ejaan yang salah.
d. *Stopword Removal*, merupakan tahapan untuk membuang kata-kata yang tidak dibutuhkan. Dalam melakukan proses ini diperlukanya algoritma *stoplist* (menghapus kata yang tidak penting) atau *wordlist* (menyimpan kata yang penting).
e. *Stemming*, merupakan tahapan untuk mentransformasi kata-kata yang terdapat dalam suatu dokumen menjadi sebuah kata dasar (*root word*) dengan menggunakan aturan aturan tertentu.

DOI: http://dx.doi.org/10.22441/fifo.2022.v14i1.007

f. *Tokenizing*, merupakan tahapan untuk memotong *string* input satuan kata penyusunya. Pemecahan kalimat menjadi kata-kata tunggal dilakukan dengan menganalisis kalimat dengan kalimat pemisah (*delimiter*) *whitespace* (spasi, tab, dan *newline*).

Setelah pembersihan data dilakukan pembobotan kata ini bertujuan untuk mengetahui seberapa besar diperlukan suatu kata untuk mewakili sebuah kalimat, sehingga diperlukanya sebuah perhitungan dan pembobotan kata untuk mengetahui tingkat dibutuhkan suatu kata. [6]. *Term Frequency-Inverse Document Frequency* (TF-IDF) merupakan metode dalam pembobotan kata dengan mengintegrasikan tf (*term frequency)* dan idf (*inverse document frequency).*
Tahapan dalam pembobotan kata yaitu:

1. *Term Frequency* (tf)
   Term Frequency merupakan jumlah banyak kata yang muncul dalam suatu dokumen. Sedangan w_{tf} merupakan bobot dari tf yang dihitung dengan algoritma.
   $$w_{tf_{t,d}} = \begin{cases} 1 + log_{10}(tf_{t,d}) \, , if \; tf_{t,d} > 0 \\ 0 \qquad\qquad\quad , if \; tf_{t,d} = 0 \end{cases} \qquad (4)$$
2. *Inverse Document Frequency* (idf)
   Documen frequency atau disebut juga dengan df merupakan jumlah banyak dokumen yang mengandung suatu kata. Sedangkan idf adalah bobot kebalikan dari bobot *document frequency* atau dapat dikatakan jumlah . Semakin besar idf maka kata jarang muncul dalam suatu dokumen.
   $$idf_t = log_{10}\left(\frac{N}{df_t}\right) \qquad (5)$$
3. *Term Frequency - Inverse Document Frequency* (TF-IDF)
   TF-IDF merupakan pembobotan yang didapatkan dari hasil perkalian dari pembobotan *term frequency* (w_{tf}) dengan *Inverse Document Frequency* (idf_t) dari suatu *term.*
   $$w_{t,d} = w_{tf} \; x \; idf_t \qquad (6)$$
   Keterangan:
   $tf_{t,d}$: Jumlah kemunculan *term* (t) pada dokumen (d)
   N : Jumlah dokumen teks
   df_t : Jumlah dokumen yang mengandung *term* (t)

Setelah setiap kata memiliki bobot sehingga dapat dibentuk sebuah model yang akan digunakan dalam proses klasifikasi dengan membagi data menjadi dua jenis yaitu, data latih dan data uji dengan menggunakan metode 10 - *fold cross validation*. Setelah dilakukan pembagian data maka dapat memasuki proses klasifikasi. Klasifikasi dapat dilakukan dengan beberapa tahap yaitu membentuk sebuah model, mengimplementasikan model dan mengevaluasi model [7]. Dalam membuat model suatu data diperlukanya sebuah algoritma. Salah satu algoritma dalam klasifikasi data adalah SVM (*Support Vector Machine*). *Support Vector Machine* adalah sebuah algoritma klasifikasi yang diperkenalkan oleh Boser, Guyon dan Vapnik pada tahun 1992. SVM *Support Vector Machine* juga dikenal sebagai teknik pembelajaran *machine learning* yang paling mutakhir setelah pembelajaran *Neural Network* [8]. Algoritma SVM merupakan klasifikasi yang memisahkan dua kelas dengan cara melakukan pencarian *hyperplane* terbaik dengan *margin* terbesar dari sampel yang diberikan [9]. *Hyperplane* adalah batas antara dua data yang berdekatan atau memiliki jarak terdekat tetapi berbeda kelas. Dalam pencarian *hyperplane* terbaik harus didapatkan sebuah *margin* terbesar dan mencari titik maksimalnya. *Margin* adalah jarak antara *hyperplane* dan *support vector*. *Support vector* merupakan data yang memiliki jarak terdekat dengan *hyperplane* atau dapat dikatakan data yang sulit untuk diklasifikasi[8]. Berikut merupakan rumus pencarian *hyperplane* terbaik yaitu:

$$\min \frac{1}{2} |w|^2 = \min \frac{1}{2} \left(\sqrt{{w_1}^2 + {w_2}^2}\right)^2 = \min \frac{1}{2} ({w_1}^2 + {w_2}^2) \qquad (7)$$
Dengan syarat $y_i \, (x_i \, w + b) - 1 \geq 0$.

Setelah dilakukan pemodelan data dilakukan klasifikasi dengan menggunakan model yang telah dibentuk kemudian dilakukan evaluasi. Tahap evaluasi dilakukan dengan menggunakan metode *confusion matrix*. Evaluasi dilakukan untuk mengetahui tingkat kebenaran dalam pembangunanan dan pembentukan model dengan memberikan nilai evaluasi berupa presentase dari akurasi (*accuracy*), sensitifitas (*recall*), *F-Measure*, nilai prediksi positif (*positive predictive value* atau *precision*) [10]. Kemudian dilakukan analisis hasil klasifikasi, Analisis hasil klasifikasi dilakukan dengan menampilkan visualisasi dari hasil klasifikasi yang telah dibentuk.

DOI: http://dx.doi.org/10.22441/fifo.2022.v14i1.007

III. Hasil Dan Pembahasan

Pada penelitian ini dilakukan pengumpulan data melalui Media sosial t*witter* telah menyediakan *website developer.twitter.com* yang dapat diakses untuk mendapatkan sebuah API *twitter*. Website *developer.twitter.com* akan memberikan API *key*, API *secret, Access Token, Access Token Secret* untuk dapat mengakses API *twitter*. Pengambilan data dilakukan menggunakan perangkat lunak *google colab* dengan bahasa pemograman *python* dan menggunakan *library tweepy*. Sebanyak 1257 data yang telah didapat dalam rentang waktu 18-02-2021 sampai 25-02-2021 dengan menggunakan kata kunci "kelas *online*" dan "kelas daring". Berikut merupakan hasil *crawling* data *tweet*:

	Tweet	User Screen Name	Status Tweet	Tanggal
0	hm aku pun entahla kenapa madam wa & miss ...	YOURKAIY	Tweet	2021-02-25 18:07:04
1	[SEBUAH UTAS]\nPerdebatan kala kelas online Gu...	ham_ilhamhammmm	Tweet	2021-02-25 17:55:27
2	@TIMSAR Aamiinn.. soalnya besok masi ada kelas...	PearlCherry24	Tweet	2021-02-25 17:44:21
3	dah ah ngantuk,met malem semua,bobo ya besok k...	jaejaewifeu	Tweet	2021-02-25 17:42:23
4	@syayonara Aku gak pernah pacaran, pacaran yan...	Taehyungiee30_	Tweet	2021-02-25 17:36:39

Gambar 2. Hasil dari *Crawling* data pada *Twitter*

Setelah melakukan *crawling* data didapatkan *tweet* data seperti pada gambar 2 dengan atribut *twee*t, *username*, status *tweet* dan tanggal yang akan disimpan pada *file* dengan format .csv. Setelah pengumpulan data, data akan melalui proses penyaringan karena masih banyaknya data yang tidak berkaitan dengan pembelajaran daring di Indonesia salah satunya seperti iklan, sehingga didata didapat sebanyak 804 data yang digunakan untuk proses selanjutnya yaitu pelabelan data. Setelah dilakukan pelabelan data oleh tiga anatator, data yang akan digunakan sebanyak 700 data yang terdiri dari 350 negatif dan 350 positif. Hasil pelabelan pada data *tweet* dapat disebut dengan kesepakatan tiga anatator. Kesepakatan yang telah dilakukan akan diukur dengan *fleiss kappa* dengan menggunakan rumus (1), (2) dan (3). Berikut merupakan contoh perhitungan *fleiss kappa* pada pelabelan data oleh tiga anatator dengan menggunakan 700 data dokumen:

Tabel 2. Data sampel perhitungan fleiss kappa

Tweet	Anatator 1	Anatator 2	Anatator 3	positif	negatif
lagi kelas online Apeni apdet?? Wah🤡🤡🤡	positif	positif	positif	3	0
...kembali ke sistem belajar online (Daring) apakah selamanya akan seperti ini? Apalagi, anak yg baru memasuki kelas pertama di Sekolah Dasar pasti dia merasa bosan dengan belajar seperti ini. Kalo di tanya juga, pasti ingin bermain dengan teman sebayanya.. 😕	negatif	negatif	negatif	0	3
//. bisa2nya lg daring trs gw lg scroll tiktok, trs ada sound yang "kenapa semua menangis, biasalah" trs gw ikut nyanyiin dan ternyata mic gw nyala.....malu bgt anjr mana satu kelas ngetawain gw😉	negatif	positif	negatif	1	2

DOI: http://dx.doi.org/10.22441/fifo.2022.v14i1.007

Pada tabel 2 merupakan data sampel dokumen yang akan digunakan untuk perhitungan *fleiss kappa*. Berikut merupakan perhitungan *fleiss kappa*:
Perhitungan p_α dapat dilihat pada bab sebelumnya dengan rumus (2) dengan banyak data n = 700 dan menggunakan sebanyak m = 3 (anatator).

$$p_\alpha = \frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{j=1}^{k}x_{ij}^2 - m}{m(m-1)}$$

$$p_\alpha = \frac{1}{700}\left(\frac{\left((3^2+0^2)-3\right)}{3(3-1)} + \frac{\left((0^2+3^2)-3\right)}{3(3-1)} + \frac{\left((1^2+2^2)-3\right)}{3(3-1)} + \cdots\right)$$

$$p_\alpha = \frac{1}{700}\left(\frac{6}{6}+\frac{6}{6}+\frac{2}{6}+\cdots\right)$$

$$p_\alpha = \frac{1}{700}(1+1+0{,}333333+\cdots)$$

$$p_\alpha = \frac{1}{700}(628)$$

$$p_\alpha = \frac{628}{700} = 0{,}897142$$

Setelah perhitungan p_α, dilakukan perhitungan p_ε dapat dilihat pada bab sebelumnya dengan rumus (3) dimana $q_j = \frac{1}{nm}\sum_{i=1}^{n}x_{ij}$:

$$q_{positif} = \frac{(3+0+1+3+3+\cdots..)}{(700)(3)} = \frac{1056}{2100} = 0{,}502857$$

$$q_{negatif} = \frac{(0+3+2+0+0\ldots.)}{(700)(3)} = \frac{1044}{2100} = 0{,}497143$$

$$p_\varepsilon = \sum_{j=1}^{k} q_j^2 = \left(q_{positif}\right)^2 + \left(q_{negatif}\right)^2$$

$$p_\varepsilon = (0{,}502857)^2 + (0{,}497143)^2 = (0{,}252865) + (0{,}247151)$$

$$p_\varepsilon = 0{,}500016$$

Setelah perhitungan p_α dan perhitungan p_ε sehingga dapat dilakukan perhitungan nilai kappa dengan menggunakan rumus (1) dengan nilai p_α sebesar 0,897143 dan p_ε sebesar 0,500016.

$$kappa = \frac{p_\alpha - p_\varepsilon}{1-p_\varepsilon} = \frac{0{,}897143 - 0{,}500016}{1 - 0{,}500016} = \frac{0{,}397127}{0{,}499984} = 0{,}794279$$

Hasil perhitungan dengan 700 data tweet didapat nilai p_α sebesar 0,897143 dan p_ε sebesar 0,500016. Hasil nilai *fleiss kappa* sebesar 0,79427 sehingga masuk dalam kategori *subtantial perfect agreement* seperti yang sudah dijelaskan dibab sebelumnya pada tabel 1 yang memiliki arti bahwa kesepakatan yang digunakan dalam pelabelan data sudah baik. Setelah dilakukan pelabelan data dilakukan pembersihan data melalui *pre-processing.* Berikut merupakan salah satu data *tweet* yang sebelum dan sesudah dilakukan *pre-processing:*

Tabel 3. Contoh Hasil Pre-Processing

Sebelum Pre-processing	Sesudah pre-processing
@SHidayat2428 @Kemdikbud_RI Gila emang nih. Masa dlm setahun ini gak bisa membuat sistem pembelajaran yg efektif dan efisien pd masa pandemi? Belajar daring tp masih pake metode belajar dlm kelas, setiap hari, dan sama waktunya dgn sekolah biasa, gila bener.	['gila', 'tahun', 'sistem', 'ajar', 'efektif', 'efisien', 'pandemi', 'ajar', 'daring', 'pakai', 'metode', 'ajar', 'kelas', 'sekolah', 'gila']

Tabel 3 merupakan contoh perubahan setelah dilakukan *pre-processing. Pre-processing* dilakukan dengan enam langkah yaitu, *case folding, data cleaning, normalization, removal stopword, stemming* dan *tokenizing*. Terdapat sebanyak 1147 kata dalam kamus *normalization*. Pada proses *removal stopword* dengan menggunakan library sastrawi terdapat sebanyak 804 kata. Kemudian proses stemiing dilakukan menggunakan library sastrawi.

Setelah melalui *pre-processing* dilakukan pembobotan kata dengan menggunakan TF-IDF. Pada penelitian ini pembobotan TF-IDF akan dilakukan oleh 700 dokumen/tweet dengan menggunakan rumus 4,5 dan 6. Berikut Perhitungan tf, idf dan tf-idf pada sampel yang ada pada tabel 2 :

Tabel 4. Perhitungan TF dan IDF pada sampel data

Kata	$Tf_{t,d}$					DF	IDF
	D1	D2	D3	D4	D5		
ajar	0	0.133333	0	0	0	1	2.09861
anak	0	0.066667	0	0	0	1	2.09861
anjing	0	0	0.076923	0	0	1	2.09861
ara	0	0	0	0.083333	0	1	2.09861
baya	0	0.066667	0	0	0	1	2.09861
biasa	0	0	0.076923	0	0	1	2.09861
bingung	0	0	0	0	0.125	1	2.09861
bosan	0	0.066667	0	0	0	1	2.09861
botcamp	0	0	0	0	0.125	1	2.09861
daring	0.25	0.133333	0.076923	0.083333	0.125	5	1
dasar	0	0.066667	0	0	0	1	2.09861
hari	0	0	0	0.083333	0	1	2.09861
hm	0	0	0	0.083333	0	1	2.09861
kagum	0.25	0	0	0	0	1	2.09861
kakak	0	0	0	0	0.125	1	2.09861
kelas	0.25	0.066667	0.076923	0.083333	0.125	5	1
kemarin	0	0	0	0	0.125	1	2.09861
kesal	0	0.066667	0	0	0	1	2.09861
kini	0.25	0	0	0	0	1	2.09861
lemas	0	0	0	0.083333	0	1	2.09861
logika	0	0	0	0	0.125	1	2.09861
lupa	0	0	0	0.083333	0	1	2.09861
main	0	0.066667	0	0	0	1	2.09861
makan	0	0	0	0.083333	0	1	2.09861
malu	0	0	0.076923	0	0	1	2.09861
menang	0	0	0.076923	0	0	1	2.09861
mic	0	0	0.076923	0	0	1	2.09861
moga	0	0	0	0.083333	0	1	2.09861
nendang	0	0	0	0	0.125	1	2.09861
ngetawain	0	0	0.076923	0	0	1	2.09861
nyala	0	0	0.076923	0	0	1	2.09861
nyanyin	0	0	0.076923	0	0	1	2.09861
pasuk	0	0.066667	0	0	0	1	2.09861
pokok	0	0	0	0.083333	0	1	2.09861
scrol	0	0	0.076923	0	0	1	2.09861
sekolah	0	0.066667	0	0	0	1	2.09861
semangat	0	0	0	0.083333	0	1	2.09861
senang	0	0	0	0.083333	0	1	2.09861
sistem	0	0.066667	0	0	0	1	2.09861
suara	0	0	0.076923	0	0	1	2.09861
teman	0	0.066667	0	0	0	1	2.09861
tiktok	0	0	0.076923	0	0	1	2.09861

DOI: http://dx.doi.org/10.22441/fifo.2022.v14i1.007

Tabel 4 merupakan hasil perhitungan tf dan idf pada sampel data lima dokumen sampel data yang telah disimpulkan dalam bentuk pada tabel. Berikut merupakan salah satu contoh perhitungan nilai tf dan idf pada D2 dengan kata "sistem":

Perhitungan TF (*Term Frequency*) dilakukan dengan menghitung banyak jumlah murni kata dalam suatu dokumen.

$$TF = \frac{jumlah\ banyak\ kata\ dalam\ suatu\ dokumen}{jumlah\ kata\ pada\ dokumen}$$

$TF_{(sistem,D2)} = \frac{1}{15} = 0.06666667$

Perhitungan IDF (*Invers Document Frequency*) dilakukan dengan menggunakan Rumus (5) dengan menjumlahkan angka 1. Hal ini dilakukan untuk mencegah pembagian nol (0) karena berapapun nilai tf jika df memiliki nilai 1 yang menghasilkan log 1 = 0 .

$idf_t = log_{10}\left(\frac{N}{df_t}\right)$ menjadi $idf_t = log_{10}\left(\frac{1+N}{1+df_t}\right) + 1$

$idf_{sistem} = log_{10}\left(\frac{6}{2}\right) + 1 = 1.098612289 + 1 = 2.098612289$

Hasil perhitungan pada tabel 4 didapatkan banyak kata dalam 5 dokumen yang digunakan sebanyak 42 kata. Setelah didapat hasil perhitungan tf dan idf maka dapat dilakukan pembobotan tf-idf dengan mengkalikan tf dan idf dengan menggunkan rumus (6):

$w_{t,d} = w_{tf}\ x\ idf_t$

$w_{sistem,D2} = 0.06666667\ x\ 2.098612289 = 0.139907$

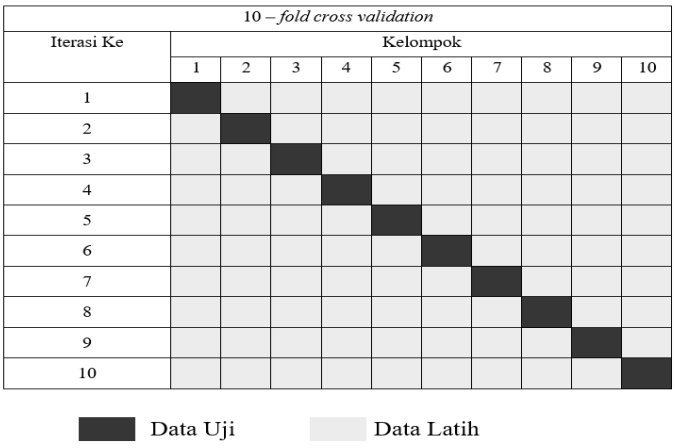
Berikut merupakan hasil pembobotan tf-idf pada sampel data dengan menggunakan lima dokumen sampel yang akan ditulisan dalam bentuk tabel, sebagai berikut :

Tabel 5. Hasil pembobotan tf-idf pada sampel

Kata	$W_{t,d}$				
	D1	D2	D3	D4	D5
ajar	0	0.279815	0	0	0
anak	0	0.139907	0	0	0
anjing	0	0	0.161432	0	0
ara	0	0	0	0.174884	0
baya	0	0.139907	0	0	0
biasa	0	0	0.161432	0	0
bingung	0	0	0	0	0.262327
bosan	0	0.139907	0	0	0
botcamp	0	0	0	0	0.262327
daring	0.25	0.133333	0.076923	0.083333	0.125
dasar	0	0.139907	0	0	0
hari	0	0	0	0.174884	0
hm	0	0	0	0.174884	0
kagum	0.25	0	0	0	0
kakak	0	0	0	0	0.262327
kelas	0.25	0.066667	0.076923	0.083333	0.125
kemarin	0	0	0	0	0.262327
kesal	0	0.139907	0	0	0
kini	0.25	0	0	0	0
lemas	0	0	0	0.174884	0
logika	0	0	0	0	0.262327
lupa	0	0	0	0.174884	0

main	0	0.139907	0	0	0
makan	0	0	0	0.174884	0
malu	0	0	0.161432	0	0
menang	0	0	0.161432	0	0
mic	0	0	0.161432	0	0
moga	0	0	0	0.174884	0
nendang	0	0	0	0	0.262327
ngetawain	0	0	0.161432	0	0
nyala	0	0	0.161432	0	0
nyanyin	0	0	0.161432	0	0
pasuk	0	0.139907	0	0	0
pokok	0	0	0	0.174884	0
scrol	0	0	0.161432	0	0
sekolah	0	0.139907	0	0	0
semangat	0	0	0	0.174884	0
senang	0	0	0	0.174884	0
sistem	0	0.139907	0	0	0
suara	0	0	0.161432	0	0
teman	0	0.139907	0	0	0
tiktok	0	0	0.161432	0	0

Pada tabel 5 menunjukan hasil akhir pembobotan tf-idf dengan menggunakan lima sampel dokumen. Pada penelitian ini pembobotan tf-idf akan dilakukan oleh 700 dokumen/tweet. Setelah dilakukan proses perhitungan tf-idf pada 700 dokumen/*tweets* yang menghasilkan *term* sebanyak 1882 *term*/kata. Setiap *term* dalam dokumen akan dijadikan sebagai fitur yang akan digunakan dalam proses klasifikasi data. Setelah proses pembobotan kata, akan dilakukan pembagian data menggunakan *10-fold cross validation* dengan menggunakan 700 dokumen dimana 700 dokumen akan dibagi menjadi 10 kelompok, sehingga setiap kelompok akan terdiri dari 70 dokumen.



Gambar 3. Ilustrasi pembagian data 10-*fold cross validation*

Pada gambar 3. dimisalkan pada iterasi pertama kelompok 1 akan dijadikan sebagai data uji dengan jumlah data sebanyak 70 dokumen/*tweet* dan pada kelompok 2,3,4,5,6,7,8,9, dan 10 akan dijadikan data latih dengan jumlah dokumen yang sama pada setiap kelompoknya sebanyak 70 dokumen/*tweet* sehingga data latih berjumlah 630 dokumen pada setiap iterasi. Seperti pada gambar 4.8 setiap iterasi yang dilakukan memiliki data uji yang bebeda, sehingga dengan metode 10-*fold cross validation* setiap kelompok atau dokumen akan dijadikan sebagai data uji. Setelah data dibagi kemudian dilakukan pelatihan data pada data latih untuk mendapatkan sebuah bentuk model klasifikasi. Setelah data dibagi kemudian dilakukan pelatihan data pada data latih untuk mendapatkan sebuah bentuk model klasifikasi. Dalam penelitian ini pemodelan data dilakukan dengan menggunakan algoritma SVM (*Support Vector Machine*). Berikut merupakan sampel data yang digunakan untuk contoh proses pemodelan data:

DOI: http://dx.doi.org/10.22441/fifo.2022.v14i1.007

Tabel 6. Data sampel pemodela data

Dokumen	kelas	daring	kini	kagum	kelas
d1	0,25	0,25	0,25	0,25	positif

Pada tabel 6. diatas merupakan sampel hasil yang telah melalui proses pembobotan tf idf yang akan dijadikan untuk contoh pencarian garis h*yperplane* terbaik, untuk dilakukan perhitungan pada tabel 6 akan dimisalkan menjadi sebagai berikut :

Tabel 7. Data Sampel SVM

x1	x2	x3	x4	kelas (y)
0,25	0,25	0,25	0,25	1

Pada tabel 7. karena data hanya terdiri dari empat fitur yaitu (x_1, x_2, x_3, x_4) maka (w) juga akan memiliki empat fitur yaitu (w_1, w_2, w_3, w_4). Dengan data tersebut akan meminimalkan margin dengan menggunakan rumus (7), yaitu:

$$\min \frac{1}{2} |w|^2 = \min \frac{1}{2} \left(\sqrt{{w_1}^2 + {w_2}^2}\right)^2 = \min \frac{1}{2} ({w_1}^2 + {w_2}^2)$$

Dengan syarat,

$$y_i (x_i w + b) - 1 \geq 0 \qquad i = 1,2,3,4, \ldots, N$$
$$y_i (x_1 w_1 + x_2 w_2 + \cdots + x_N w_N + b) - 1 \geq 0$$

sehingga didapatkan satu persamaan pada contoh sampel tabel 4.8, yaitu :

$$(0{,}25w_1 + 0{,}25w_2 + 0{,}25w_3 + 0{,}25w_4 + b) \geq 1 ,$$
$$\text{untuk } y_1 = 1;\ x_1, x_2, x_3, x_4 = 0{,}25$$

Pada contoh diatas hanya menggunakan satu dokumen sehingga hanya didapatkan satu persamaan. Pada penelitian ini digunakan sebanyak 700 data yang nantinya untuk setiap dokumen akan dilakukan perhitung satu persatu seperti pada contoh sampel tersebut. Semakin banyak dokumen akan semakin banyak fitur yang akan digunakan. Setelah didapatkan beberapa persamaan sehingga dilakukan eliminasi satu persatu sehingga didapat nilai w_1, w_2, w_3, w_4. Setelah itu dilakukan subtitusi pada persamaan $x_1 w_1 + x_2 w_2 + \cdots + x_N w_N + b = 0$, Sehingga didapatkan persamaan garis *linear* yang akan menunjukan titik koordinat yang membentuk garis persamaan *hyperplane* sebagai fungsi klasifikasi. Pencarian garis *hyperlane* terbaik dilakukan dengan menimumkan margin sehingga dilakukan pencarian dengan menggunakan rumus (7) dengan mengambil nilai yang paling minimum. Model yang telah dibentuk akan diuji untuk mengetahui kemampuan model dengan mengklasifikasi data uji.

Sebelum melakukan pemodelan data dilakukan *Hyperparameter Tunning* dengan menggunakan *Grid Search CV*. Empat kernel yang diujikan yaitu *Linear*, RBF, *Sigmoid* dan *Polynomial*. Uji parameter dilakukan dengan meggunakan parameter c yaitu 1, 10, 100, 1000 dan parameter gamma yaitu 0.01, 0.1, 1, 10, 100. Uji ini dilakukan dengan menggunakan pebagian data 80:20. Setelah dilakukan uji untuk setiap parameter didapatkan hasil parameter yang terbaik pada setiap kernel. Berikut merupakan hasil parameter yang terbaik pada setiap kernel :

Tabel 6. Hasil Parameter Terbaik *Grid Search CV* setiap kernel

Parameter Terbaik	Kernel			
	Linear	RBF	Poly	Sigmoid
c	1	10	0,1	1000
gamma	0,1	0,1	10	0,01

Tabel 6 menujukan hasil parameter terbaik *Grid Search CV* untuk setiap kernel. Setelah didapatkan parameter terbaik untuk setiap kernel, selanjutnya akan dilakukan implementasi dengan menggunakan parameter terbaik pada tabel 4 untuk setiap kernel. Berikut merupakan hasil akurasi untuk setiap kernel :

Tabel 7. Hasil akurasi parameter terbaik *Grid Search CV*

	Kernel			
	Linear	*RBF*	*Poly*	*Sigmoid*
akurasi	71,4%	70,0%	63,6%	68,5%

Pada tabel 7. nilai akurasi yang lebih tinggi dari metode kernel lainya adalah dengan menggunakan kernel *linear* dengan hasil akurasi sebesar 71,4% pada parameter c = 1 dan gamma = 0,1. Oleh karena itu, penelitian ini akan menggunakan kernel *linear* sebagai parameter untuk pemodelan data pada proses klasifikasi pada data tweet mengenai pembelajaran daring di Indonesia. Dalam penelitian ini pemodelan data dilakukan dengan menggunakan algoritma SVM (*Support Vector Machine*). Pemodelan ini dilakukan dengan menggunakan beberapa *library* pada *python*.

Tabel 8. Hasil Evaluasi

Iterasi ke -	Positif			Negatif			*Accuracy*
	Precision	*Recall*	*F1-Score*	*Precision*	*Recall*	*F1-Score*	
1	**0,80**	**0,85**	**0,82**	**0,86**	**0,85**	**0,83**	**0,83**
2	0,89	0,70	0,78	0,63	0,85	0,72	0,76
3	0,74	0,76	0,75	0,77	0,75	0,76	0,76
4	0,66	0,82	0,73	0,86	0,71	0,78	0,76
5	**0,54**	**0,54**	**0,54**	**0,54**	**0,54**	**0,54**	**0,54**
6	0,66	0,72	0,69	0,74	0,68	0,71	0,70
7	0,83	0,81	0,82	0,80	0,82	0,80	0,81
8	0,71	0,66	0,68	0,63	0,71	0,66	0,67
9	0,51	0,60	0,55	0,66	0,57	0,60	0,59
10	0,69	0,80	0,74	0,83	0,72	0,77	0,76

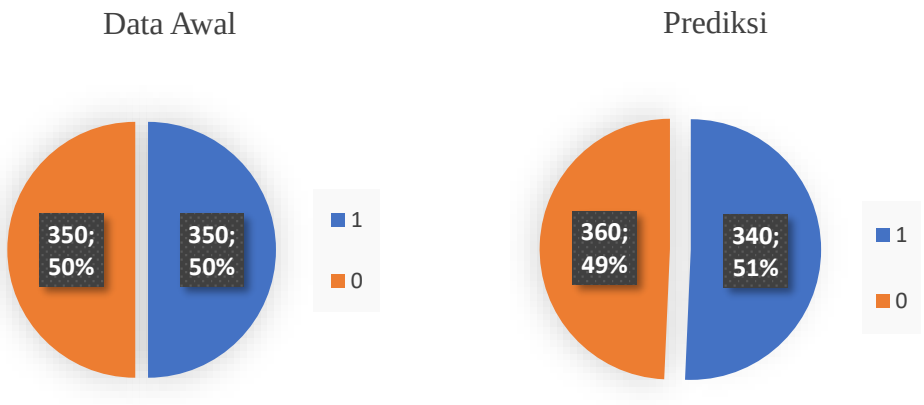
Tabel 8. Merupakan hasil dari evaluasi data *confussion matrix* pada setiap iterasi dengan 10-*fold cross validation* menggunakan SVM (*Support Vector Machine*). Dari hasil perhitungan pada tabel 8 dapat dilihat bahwa pada iterasi ke-1 memiliki nilai akurasi tertinggi yang bernilai 0,83 atau 83%. Sedangkan pada iterasi ke-5 memiliki nilai akurasi terendah dengan nilai akurasi sebesar 0,54 atau 54%. Pada tabel 8 dapat dihitung rata-rata dari *Precission*, *Recall*, *F1-Score* dan *Accuracy*. Berikut merupakam hasil dari rata-rata evaluasi data pada *confussion matrix*:

Tabel 9. Hasil rata-rata evaluasi

	Positif			Negatif			*accuracy*
	precision	*recall*	*f1-score*	*precision*	*recall*	*f1-score*	
Rata-Rata	0,703	0,726	0,712	0,732	0,720	0,717	0,717143

Berdasarkan tabel 9. Pada penelitian ini didapatkakan rata-rata hasil evaluasi pada tabel 4 dengan *accuracy* sebesar 0,72 atau 72%, nilai *precission* positif sebesar 0,70 atau 70% dan *precission* negatif sebesar 0,73 atau 73% , *recall* positif sebesar 0,73 atau 73% dan *recall* negatif sebesar 0,72 atau 72%, *f1-score* positif sebesar 0,71 atau 71% dan *f1-score* negatif sebesar 0,72 atau 72%. Berikut merupakan hasil klasifikasi data sebelum dan setelah klasifikasi:

DOI: http://dx.doi.org/10.22441/fifo.2022.v14i1.007



Gambar 4. Grafik data sebelum dan setelah klasifikasi

Pada gambar 4. terlihat grafik dengan bagian biru "1" bernilai "positif" dan bagian jingga dengan "0" opini yang bernilai "negatif" atau jumlah perubahan data setelah dan sebelum dilakukan klasifikasi data dengan pemodelan menggunakan algoritma SVM. Setelah dilakukan klasifikasi menghasilkan prediksi data sebesar 340 berlabel sentimen positif dan 360 berlabel sentimen negatif. Berikut merupakan hasil *confusion matrix* sebelum dan setelah klasifikasi :

Tabel 10. Hasil *confusion matrix*

	True Class	
Predicted Class	*Positive*	*Negative*
Positive	TP : 239	FP : 101
Negative	FN : 111	TN : 249

Pada tabel 10. merupakan hasil confusion matrix pada label data sebeum dilakukan klasifikasi dengan data hasil prediksi setelah dilakukan klasifikasi. Pada tabel 10 terlihat bahwa data yang diklasifikasikan benar sebanyak 488 data dan data yang salah terklasifikasi salah sebanyak 212.

IV. Kesimpulan

Pada penelitian ini dapat ditarik kesimpulan, proses klasifikasi dilakukan dengan pelabelan data secara manual pada 700 data tweet yang terdiri dari 350 data berlabel positif dan 350 data berlabel negatif. Kemudian dilakukan pembersihan data yaitu, *pre-processing* (*case folding, data cleaning, normalization, removal stopword, stemming, tokenizing)*. Setiap kata akan diberikan bobot dengan menggunakan TF-IDF, kemudian dilakukan pembagian data dengan menggunakan *10-fold cross validation* dan pemodelan data SVM. Hasil yang didapat pada penelitian ini sebesar, iterasi ke-1 memiliki nilai akurasi tertinggi yang bernilai 0,83 atau 83%. Sedangkan pada iterasi ke-5 memiliki nilai akurasi terendah dengan nilai akurasi sebesar 0,54 atau 54%. Dengan hasil evaluasi didapatkakan rata-rata hasil *accuracy* sebesar 0,72 atau 72% dengan nilai *precission* positif sebesar 0,70 atau 70% dan *precission* negatif sebesar 0,73 atau 73% , *recall* positif sebesar 0,73 atau 73% dan *recall* negatif sebesar 0,72 atau 72%, *f1-score* positif sebesar 0,71 atau 71% dan *f1-score* negatif sebesar 0,72 atau 72%. Hasil klasifikasi data *tweet* menjadi 340 data dengan sentimen positif dan 360 data sentimen negatif.

Daftar Pustaka

[1] K. P. dan Kebudayaan, "Kemendikbud Terbitkan Pedoman Penyelenggaraan Belajar dari Rumah, Surat Edaran Nomor 4 Tahun 2020," 2020. https://www.kemdikbud.go.id/main/blog/2020/05/kemendikbud-terbitkan-pedoman-penyelenggaraan-belajar-dari-rumah.

[2] F. A. Nugraha, N. H. Harai, and R. Habibi, *Analisis Sentimen Terhadap Pembatasan Sosial Menggunakan Deep Learning*. Bandung: Kreatif Industri Nusantara, 2020.
[3] S. Fauziah, D. N. Sulistyowati, and T. Asra, *Optimasi Algoritma Vector Space Model Dengan Algoritma K-Nearest Neighbour Pada Pencarian Judul Artikel Jurnal*, vol. 15, no. 1. 2019.
[4] T. R. Nichols, P. M. Wisner, G. Cripe, and L. Gulabchand, "Putting the kappa statistic to use," *Qual. Assur. J.*, vol. 13, no. 3–4, pp. 57–61, 2010, doi: 10.1002/qaj.481.
[5] W. A. Luqyana, I. Cholissodin, and R. S. Perdana, "Analisis Sentimen Cyberbullying Pada Komentar Instagram dengan Metode Klasifikasi Support Vector Machine," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 2, no. 11, pp. 4704–4713, 2018.
[6] P. P. A. Arsya Monica Pravina, Imam Cholissodin, "Analisis Sentimen Tentang Opini Maskapai Penerbangan pada Dokumen Twitter Menggunakan Algoritme Support Vector Machine (SVM)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 3, pp. 2789–2797, 2019, [Online]. Available: http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/4793.
[7] D. A. Nasution, H. H. Khotimah, and N. Chamidah, "Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN," *Comput. Eng. Sci. Syst. J.*, vol. 4, no. 1, p. 78, 2019, doi: 10.24114/cess.v4i1.11458.
[8] D. Yosmita Praptiwi, "Analisis Sentimen Online Review Pengguna E-Commerce Menggunakan Metode Support Vector Machine Dan Maximum Entropy," 2018.
[9] Chapman and Hall/CRC, *The Top Ten Algorithms in Data Mining*. Taylot & francis Group, LLC, 2009.
[10] B. Santoso, A. I. S. Azis, and Zohrahayaty, *Machine Learning & Reasoning Fuzzy Logic Algoritma, Manual, Matlab, & Rapid Miner*. Yogyakarta: Deepublish, 2020.