ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 11, Nomor 3, Desember 2022

SENTIMENT ANALYSIS OF ONLINE LECTURES USING K-NEAREST NEIGHBORS BASED ON FEATURE SELECTION

Junadhi[1], Agustin[2], Mi'rajul Rifqi[3], M. Khairul Anam[4]

[1]Teknik Informatika, STMIK Amik Riau
[2,4]Teknologi Informasi, STMIK Amik Riau
[3]Sistem Informasi, Universitas Pasir Pengaraian

email: junadhi@sar.ac.id[1], agustin@sar.ac.id[2], mirajulrifqi@gmail.com[3], khairulanam@sar.ac.id[4]

Abstract

Online lecture is a distance learning system that utilizes information technology in its implementation. Although it has been agreed, this lecture system has caused controversy. Not infrequently online lectures are considered to bring a variety of new obstacles in lectures, and not a few also consider that online lectures are the most appropriate solution to continue to run lecture activities in the midst of alarming pandemic conditions. In response to this policy, many people expressed various kinds of opinions and views on the implementation of online lectures which are generally stated on social media, one of which is through Twitter. Sentiment analysis is a branch of the science of machine learning that is carried out to obtain useful information or new knowledge by extracting, understanding, and processing text data automatically. Several methods are widely used by researchers to classify sentiment analysis datasets including K-Nearest Neighbor (K-NN). K-NN will be adapted to classify online lecture datasets because K-NN can produce good accuracy on a large amount of data. The presence of feature selection also helps machine learning in improving its performance. The purpose of this study was to determine student sentiment toward online lectures and to determine the level of accuracy of the combination of K-NN with various feature selections. Based on 780 tweets data, a classification of 394 positive sentiments, 320 negative sentiments, and 39 neutral sentiments was obtained. So, the results of student opinions are on POSITIVE sentiments. The accuracy result of the K-NN Algorithm was 56% and the accuracy of the K-NN Algorithm + Forward Selection was 51%, the accuracy of the KNN Algorithm + Adabost was 54%, and the accuracy of the KNN Algorithm + Genetic Algorithm was 55%.

Keywords: *Online lectures, Sentiment analysis, K-NN, Feature Selection, Tweets*

Diterima Redaksi: 19-08-2022 | **Selesai Revisi:** 11-12-2022 | **Diterbitkan Online:** 27-12-2022
DOI: https://doi.org/10.23887/janapati.v11i3.51531

INTRODUCTION

Online learning is the application of *online* distance education [1]. Lectures or online learning are an educational innovation that involves elements of information technology in learning. According to [2] online learning is a distance education system with a set of teaching methods where there are teaching activities that are carried out separately from learning activities. This learning aims to increase access for students to obtain better and quality learning because with online learning, it will provide opportunities for students to be able to take part in a certain lesson or course. Online learning is considered to be the best solution to teaching and learning activities in the midst of the COVID-19 pandemic. The change in the learning process in higher education or lectures from face-to-face to online is a decision that must be made so that educational goals can be implemented effectively and efficiently. Although it has been agreed, this lecture system has caused controversy.

The emergence of new policies is certainly not an easy thing to follow. The community, especially students, lecturers, and the academic community who are directly involved with lecture activities, need to adapt to this policy. Not infrequently online lectures are considered to bring a variety of new obstacles in lectures, and not a few also consider online lectures as the most appropriate solution to continue to carry out lecture activities in the midst of an alarming pandemic condition. In response to this policy, many people expressed various kinds of opinions, opinions, and views on the implementation of online lectures. These opinions are generally expressed on social media, one of which is through Twitter. Twitter became a popular social networking site used today. The public can easily express a wide variety of their comments, thoughts, and responses related to the current conditions on social media twitter. Based on a report published by the Indonesian Internet Service Providers Association (APJII), 1.7% of the total number of internet users in Indonesia or around 291,000,417 out of 171,176,716.8 people is active users of social media twitter.

Twitter is a fairly good medium in obtaining data because the level of accuracy of opinion sentences (tweets) uploaded to twitter is considered quite high if it is used to find out how people think about a topic [3]. Through twitter, users can upload content as they wish [4]. The content is in the form of opinions, sentiments, or emoticons, which can be data to analyze a certain trend or topic [5]. Efforts to analyze the data are called sentiment analysis or opinion mining [6]. Sentiment analysis is a branch of science of machine learning that is carried out to obtain useful information or new knowledge by extracting, understanding, and processing text data automatically [7]. Through the process of sentiment analysis, it will be seen how the tendency of one's opinion towards a topic or problem is by establishing the classification of sentiment into two or more classes [8].

Machine Learning has been widely used to assist in supporting stakeholders' decisions in making policies as done by [9]–[11]. Several methods are widely used by researchers to classify sentiment analysis datasets including K-NN with an *accuracy* of 87.00%, *Support Vector Machine* with an accuracy of 86.00%, and *Supervised Machine Learning* with an *accuracy* of 87.00%. Online lecture datasets are used as training data and test data. K-NN will be adapted to classify online lecture datasets because K-NN can produce good accuracy on a large amount of data and K-NN is a straightforward and powerful algorithm, so this study used this method [12].

The presence of *feature selection* also helps *machine learning* in improving its performance. Feature selection algorithms that are widely used for classification cases include *Forward Selection, Adaboost* and *Genetic Algorithms* [13]–[16]. In this study, several *feature selections* will be applied including *Forward Selection, Adaboost* and *Genetic Algorithms* will be implemented to select relevant features, so that the accuracy of K-NN can be maximized. The purpose of this study was to determine student sentiment towards online lectures on social media twitter (positive, negative and neutral) and find out the level of accuracy of the combination of K-NN with various feature selection.

The use of feature selection to improve the classification accuracy of *machine learning* has been widely used by researchers, as done by [17] who applied the *feature selection* method to improve breast cancer diagnosis results. The breast cancer dataset used was the Wisconsin Breast Cancer Database (WBCD) dataset. The feature selection method used was F-Score, and the classification algorithms used were SMO, Naïve Bayes, *Multilayer Perceptron*, and C4.5. The evaluation method used was 10 *fold cross validation*. The results showed that the highest accuracy was obtained by the combination of F-score with Naïve Bayes, which was 97.65% with the number of attributes used 6 out of 9 attributes. Meanwhile, the different test tests also showed that the use of F-Score in naive bayes had significant differences. Research [18] uses the *weight by correlation* selection feature in the Support Vector Machine algorithm for Election Commission Sentiment Analysis. The dataset used was obtained from Twitter with a *crawling* technique with the keyword KPU. The evaluation method used 10 fold cross validation. The comparison of accuracy results showed that the accuracy of weight by correlation and support vector machine was higher than the accuracy of support vector machine without weight by correlation, which was 81.18% from 66.49%.

RESEARCH METHODOLOGY

This study used a framework consisting of several steps. The following is presented in Figure 3.1 below:

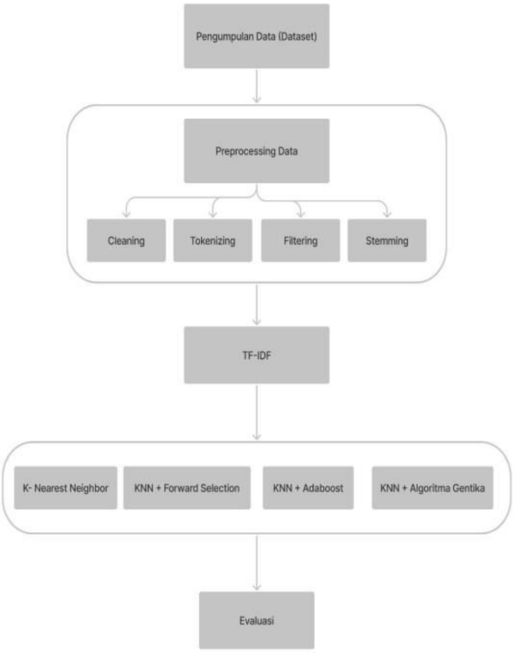
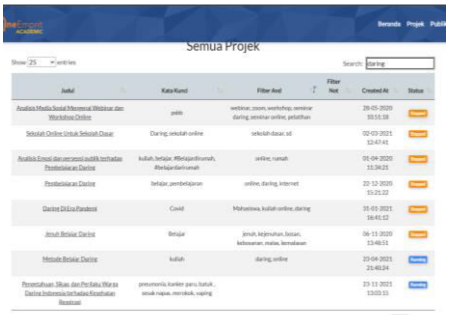


Figure 1. Research Methodology

In the research methodology above, each stage process can be explained as follows:

A. Data Collection (Dataset)

The author took the dataset on the emprit drone application. The dataset used was 780 tweets. The dataset obtained started from June 2021 to August 2021. The comparison of training data and testing data was 70: 30.



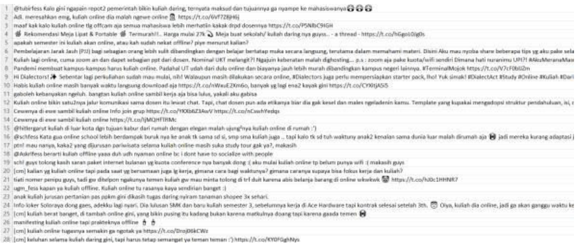
Semua Projek

Show 25 entries   Search: daring

Judul	Kata Kunci	Filter And	Filter Not	Created At	Status
Analisis Media Sosial Mengenai Webinar dan Workshop Online	pddt	webinar, zoom, workshop, seminar daring, seminar online, pelatihan		28-05-2020 10:51:18	[illegible]
Sekolah Online Untuk Sekolah Dasar	Daring, sekolah online	sekolah dasar, sd		02-03-2021 12:47:41	[illegible]
Analisis Emosi dan persepsi publik terhadap Pembelajaran Daring	kuliah, belajar, #Belajardirumah, #belajardarirumah	online, rumah		01-04-2020 11:34:21	[illegible]
Pembelajaran Daring	belajar, pembelajaran	online, daring, internet		22-12-2020 15:21:22	[illegible]
Daring Di Era Pandemi	Covid	Mahasiswa, kuliah online, daring		31-01-2021 16:41:12	[illegible]
Jenuh Belajar Daring	Belajar	jenuh, kejenuhan, bosan, kebosanan, malas, kemalasan		06-11-2020 13:48:51	[illegible]
Metode Belajar Daring	kuliah	daring, online		23-04-2021 21:40:24	Running
Pengetahuan, Sikap dan Perilaku Warga Daring Indonesia terhadap Kesehatan Respirasi	pneumonia, kanker paru, batuk, sesak napas, merokok, vaping			23-11-2021 13:03:15	Running

Figure 2. Emprit Drone Page

The dataset obtained from the emprit drone was then converted to .csv format. Here is the dataset.



```
1  @tubirfess Kalo gini ngapain repot2 pemerintah bikin kuliah daring, ternyata maksud dan tujuannya ga nyampe ke mahasiswanya
2  Adi. meresahkan emg, kuliah online dia malah ngewe online https://t.co/6Vf728jH6j
3  maaf kak kalo kuliah online tlg offcam aja semua mahasiswa lebih merhatiin kakak drpd dosennya https://t.co/P5NlbC9IGH
4  Rekomendasi Meja Lipat & Portable Termurah!!.. Harga mulai 27k Meja buat sekolah/ kuliah daring nya guyss.. - a thread - https://t.co/hGgo1Olg0s
5  apakah semester ini kuliah akan online, atau kah sudah nekat offline? piye menurut kalian?
6  Pembelajaran Jarak Jauh (PJJ) bagi sebagian orang lebih sulit dibandingkan dengan belajar bertatap muka secara langsung, terutama dalam memahami materi. Disini Aku mau nyoba share beberapa tips yg aku pake sela
7  Kuliah lagi online, cuma zoom an dan dapet sebagian ppt dari dosen. Nominal UKT melangit?! Ngajuin keberatan malah dighosting... p.s : zoom aja pake kuota/wifi sendiri Dimana hati nuranimu UPI?! #AkuMeranaMas
8  Pandemi membuat kampus-kampus harus kuliah online. Padahal UT udah dari dulu online dan biayanya jauh lebih murah dibandingkan kampus negeri lainnya. #TerminalMojok https://t.co/V7cF0bIJ2m
9  Hi Dialectors! Sebentar lagi perkuliahan sudah mau mulai, nih! Walaupun masih dilakukan secara online, #Dialectors juga perlu mempersiapkan starter pack, lho! Yuk simak! #DialectAct #Study #Online #Kuliah #Dari
10 Habis kuliah online masih banyak waktu langsung download aja https://t.co/nWauE2Xm6o, banyak yg lagi ena2 kayak gini https://t.co/CYXItjASi5
11 gaboleh kebanyakan ngeluh. bangtan kuliah online sambil kerja aja bisa lulus, yakali aku gabisa
12 Kuliah online bikin satu2nya jalur komunikasi sama dosen itu lewat chat. Tapi, chat dosen pun ada etikanya biar dia gak kesel dan males ngeladenin kamu. Template yang kupakai mengadopsi struktur pendahuluan, isi, c
13 Cewenya di ewe sambil kuliah online Info join grup https://t.co/YKXb6Z3AwV https://t.co/nCawhYedqs
14 Cewenya di ewe sambil kuliah online https://t.co/ljMQHfTRMc
15 @hitlergarut kuliah di luar kota dgn tujuan kabur dari rumah dengan elegan malah ujung²nya kuliah online di rumah :')
16 @schfess Kata gua online school lebih berdampak buruk nya ke anak tk sama sd si, smp sma kuliah juga .. tapi kalo tk sd tuh waktuny anak2 kenalan sama dunia luar malah dirumah aja jadi mereka kurang adaptasi j
17 ptn! mau nanya, kaka2 yang dijurusan pariwisata selama kuliah online masih suka study tour gak ya?, makasih
18 @Askrlfess berarti kuliah offline yaaa duh udh nyaman online bc i dont have to socialize with people
19 sch! guys tolong kasih saran paket internet bulanan yg kuota conference nya banyak dong :( aku mulai kuliah online tp belum punya wifi :( makasih guys
20 [cm] kalian yg kuliah online tapi pada saat yg bersamaan juga lg kerja, gimana cara bagi waktunya? gimana caranya supaya bisa fokus kerja dan kuliah?
21 tiati nomer penipu guys, tadi gw ditelpon ngakunya temen kuliah gw mau minta tolong di trf duit karena abis belanja barang di online wkwkwk https://t.co/hJ0c1HHNR7
22 ugm_fess kapan ya kuliah offline. Kuliah online tu rasanya kaya sendirian banget :)
23 anak kuliah jurusan pertanian pas pkkm gini dikasih tugas daring nyiram tanaman shopee 3x sehari.
24 Info loker Soloraya dong gaes, adekku lagi nyari. Dia lulusan SMK dan baru kuliah semester 3, sebelumnya kerja di Ace Hardware tapi kontrak selesai setelah 3th. Oiya, kuliah dia online, jadi ga akan ganggu waktu ke
25 [cm] kuliah berat banget, di tambah online gini, yang bikin pusing itu kadang bukan karena matkulnya doang tapi karena gaada temen
26 manifesting kuliah online tapi prakteknya offline
27 [cm] kuliah online tugasnya semakin ga ngotak ya https://t.co/DrojD6kCWz
28 [cm] keluhan selama kuliah daring gini, tapi harus tetap semangat ya teman teman :') https://t.co/KYDFGghNys
```

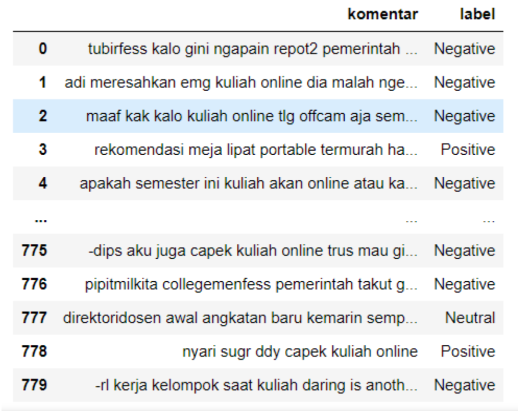
Figure 3. Online Lecture Dataset

B. Preprocessing Data

Preprocessing is one of the stages of eliminating problems that can interfere with the results of the data process [19]. Sentiment research on online lectures using text-type data went through the types of processes include cleaning, tokenizing, filtering, and stemming.

1. Cleaning

The first stage carried out was the *cleaning* stage. This stage is almost always included when doing *text preprocessing*. Because the data we have was not always structured and consistent in the use of capital letters, the role of cleaning was for the removal of punctuation, generalizing the use of capital letters, eliminating duplicate tweet data and correcting spelling [20]. Figure 4 is the result of *cleaning*.

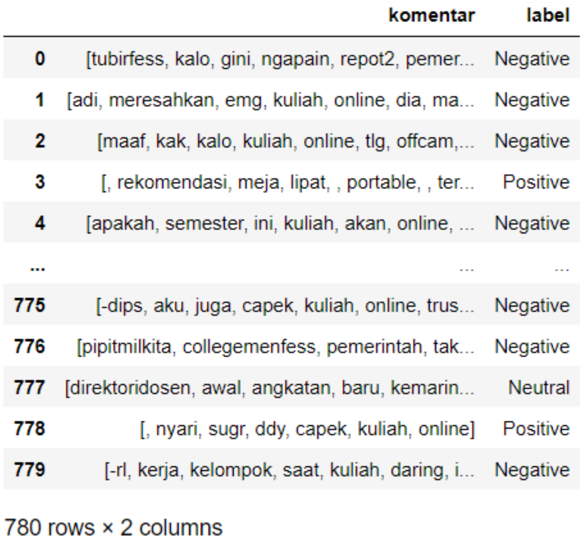


	komentar	label
0	tubirfess kalo gini ngapain repot2 pemerintah ...	Negative
1	adi meresahkan emg kuliah online dia malah nge...	Negative
2	maaf kak kalo kuliah online tlg offcam aja sem...	Negative
3	rekomendasi meja lipat portable termurah ha...	Positive
4	apakah semester ini kuliah akan online atau ka...	Negative
...	...	...
775	-dips aku juga capek kuliah online trus mau gi...	Negative
776	pipitmilkita collegemenfess pemerintah takut g...	Negative
777	direktoridosen awal angkatan baru kemarin semp...	Neutral
778	nyari sugr ddy capek kuliah online	Positive
779	-rl kerja kelompok saat kuliah daring is anoth...	Negative

Figure 4. Cleaning results

2. Tokenizing

Tweet data or datasets consisting of sentences needed a data analysis process to break sentences into words or called tokens [21]. By *tokenizing* we can distinguish word separators or not. If using the python programming language, usually *tokenizing* also includes the process of *removing numbers, removing punctuation* such as symbols and unimportant punctuation, and *removing whitespace* [22]. Figure 5 is the result of *tokenizing.*



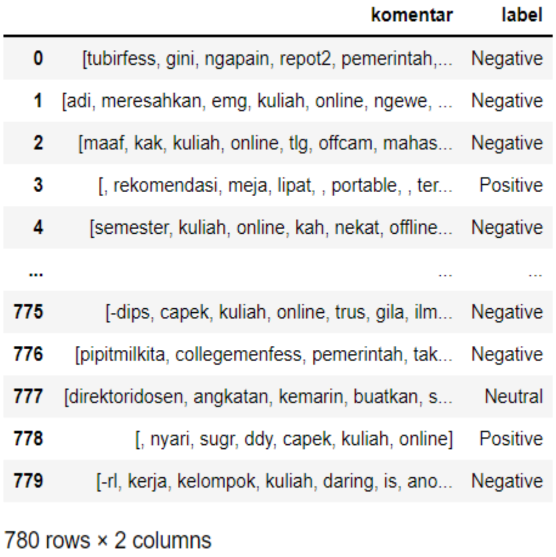
	komentar	label
0	[tubirfess, kalo, gini, ngapain, repot2, pemer...	Negative
1	[adi, meresahkan, emg, kuliah, online, dia, ma...	Negative
2	[maaf, kak, kalo, kuliah, online, tlg, offcam,...	Negative
3	[, rekomendasi, meja, lipat, , portable, , ter...	Positive
4	[apakah, semester, ini, kuliah, akan, online, ...	Negative
...	...	...
775	[-dips, aku, juga, capek, kuliah, online, trus...	Negative
776	[pipitmilkita, collegemenfess, pemerintah, tak...	Negative
777	[direktoridosen, awal, angkatan, baru, kemarin...	Neutral
778	[, nyari, sugr, ddy, capek, kuliah, online]	Positive
779	[-rl, kerja, kelompok, saat, kuliah, daring, i...	Negative

780 rows × 2 columns

Figure 5. Tokenizing Results

3. Filtering

The continuation of the tokenizing stage is the *filtering* stage which is used to retrieve important words from the token result [23]. A common word that usually appears and has no meaning is called a *stopword*. An example is the use of connecting words such as and, which, as well as, after, and others. The removal of this *stopword* can reduce the size of the index and processing time. In addition, it can also reduce noise levels, but sometimes stopping does not always increase the retrieval value [24]. The construction of a less careful *stopword* list (called a stoplist) can worsen the performance of an Information Retrieval (IR) system. There is no definite conclusion that the use of *stopping* will always increase the retrieval value because in some studies the results obtained tend to vary. Figure 6 is the result of the *filtering*.



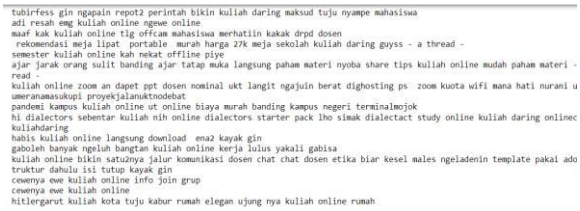
	komentar	label
0	[tubirfess, gini, ngapain, repot2, pemerintah,...	Negative
1	[adi, meresahkan, emg, kuliah, online, ngewe, ...	Negative
2	[maaf, kak, kuliah, online, tlg, offcam, mahas...	Negative
3	[, rekomendasi, meja, lipat, , portable, , ter...	Positive
4	[semester, kuliah, online, kah, nekat, offline...	Negative
...	...	...
775	[-dips, capek, kuliah, online, trus, gila, ilm...	Negative
776	[pipitmilkita, collegemenfess, pemerintah, tak...	Negative
777	[direktoridosen, angkatan, kemarin, buatkan, s...	Neutral
778	[, nyari, sugr, ddy, capek, kuliah, online]	Positive
779	[-rl, kerja, kelompok, kuliah, daring, is, ano...	Negative

780 rows × 2 columns

Figure 6. Filtering Results

4. Stemming

The *stemming* stage is a stage that is also needed to reduce the number of different indices of a single data so that a word which has both a suffix and a prefix will return to its basic form [25]. In addition, it is also to group other words having similar basic words and meanings but in different forms because they get different affixes. In the NLTK *library*, there are also modules for the *stemming* process, including *porter, lancaster, wordnet,* and *snowball.* However, again, these modules do not yet support Indonesian text. Figure 7 is the result of *stemming*.



```
tubirfess gin ngapain repot2 perintah bikin kuliah daring maksud tuju nyampe mahasiswa
adi resah emg kuliah online ngewe online
maaf kak kuliah online tlg offcam mahasiswa merhatiin kakak drpd dosen
 rekomendasi meja lipat  portable  murah harga 27k meja sekolah kuliah daring guyss - a thread -
semester kuliah online kah nekat offline piye
ajar jarak orang sulit banding ajar tatap muka langsung paham materi nyoba share tips kuliah online mudah paham materi -
read -
kuliah online zoom an dapet ppt dosen nominal ukt langit ngajuin berat dighosting ps  zoom kuota wifi mana hati nurani u
umeranamasukupi proyekjalanuktnodebat
pandemi kampus kuliah online ut online biaya murah banding kampus negeri terminalmojok
hi dialectors sebentar kuliah nih online dialectors starter pack lho simak dialectact study online kuliah daring onlinec
kuliahdaring
habis kuliah online langsung download  ena2 kayak gin
gaboleh banyak ngeluh bangtan kuliah online kerja lulus yakali gabisa
kuliah online bikin satu2nya jalur komunikasi dosen chat chat dosen etika biar kesel males ngeladenin template pakai ado
truktur dahulu isi tutup kayak gin
cewenya ewe kuliah online info join grup
cewenya ewe kuliah online
hitlergarut kuliah kota tuju kabur rumah elegan ujung nya kuliah online rumah
```

Figure 7. *Stemming* Results

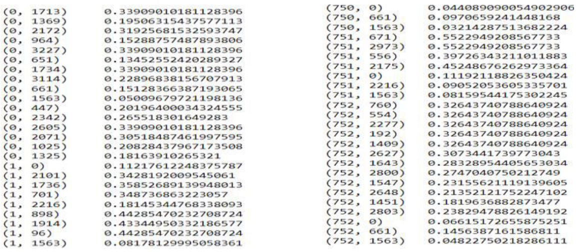
C. TF-IDF

Once the preprocessing stage was completed, the next stage was the weighting of words with the Term Frequency-Inverse Document Frequency (TF-IDF). The TF IDF method is a way of weighting the relationship of a word (term) to a document [26]. In this weighting, each word is parsed first, and then counts the appearance of each word. This method will calculate the Term Frequency (TF) and Inverse Document Frequency (IDF) values on each token (word) in each document in the corpus. This method will calculate the weight of each t token in the document with the Equation 1:

Wdt = tfdt * Idft (1)

Where:
d : the d- document
t : the t- word from the keyword
W : weight of the d- document against the t-word
tf : the number of words searched for in a document
IDF :Inversed Document Frequency
IDF values obtained from
IDF : log 2 (D/df)
where
D : total documents
df : many documents containing the searched word

After the weight (W) of each document is known, a sorting process is carried out in which the greater the value of *W* is, the greater the degree of similarity of the document to keywords will be, and vice versa [27]. Figure 8 is the result of the TF-IDF in this study.



```
(0, 1713)     0.33909010181128396
(0, 1369)     0.19506315437577113
(0, 2172)     0.31925681532593747
(0, 964)      0.15288757487893806
(0, 3227)     0.33909010181128396
(0, 651)      0.13452552420289327
(0, 1734)     0.33909010181128396
(0, 3114)     0.22896838156707913
(0, 661)      0.15128366387193065
(0, 1563)     0.05009679721198136
(0, 447)      0.20196400034324555
(0, 2342)     0.265518301649283
(0, 2605)     0.33909010181128396
(0, 2071)     0.30518487461997595
(0, 1025)     0.20828437967173508
(0, 1325)     0.18163910265321
(1, 0)        0.11217612248375787
(1, 2101)     0.3428192009545061
(1, 1736)     0.35852689139948013
(1, 701)      0.348736863223057
(1, 2216)     0.18145344768338093
(1, 898)      0.44285470232708724
(1, 1914)     0.43344950332186577
(1, 96)       0.44285470232708724
(1, 1563)     0.08178129995058361
```

```
(750, 0)      0.044089090054902906
(750, 661)    0.0970659241448168
(750, 1563)   0.03214287513682224
(751, 671)    0.5522949208567733
(751, 2973)   0.5522949208567733
(751, 556)    0.39726343211011883
(751, 2175)   0.45248676262973364
(751, 0)      0.11192118826350424
(751, 2216)   0.09052053605335701
(751, 1563)   0.08159544175302245
(752, 760)    0.32643740788640924
(752, 554)    0.32643740788640924
(752, 2277)   0.32643740788640924
(752, 192)    0.32643740788640924
(752, 1409)   0.32643740788640924
(752, 2627)   0.3073441739773043
(752, 1643)   0.28328954405653034
(752, 2800)   0.2747040750212749
(752, 1547)   0.23155621119139605
(752, 2648)   0.21352121752247102
(752, 1451)   0.1819636882873477
(752, 2803)   0.23829478826149192
(752, 0)      0.06615172655875251
(752, 661)    0.1456387161586811
(752, 1563)   0.04822750218286111
```

Figure 8. Tf-Idf Calculation Results

D. Implementation of the K-Nearest Neighbor Algorithm

The K-Nearest Neighbor algorithm according to [10] belongs to the group of instance-based learning. This algorithm is also one of the lazy learning techniques [28], [29]. K-NN is performed by looking for the group k of objects in the training data that is closest (similar) to the object in the new data or testing data [24]. A classification system is needed as a system that is able to find information. The calculation of the distance of neighborliness utilized the euclidean algorithm as shown in the Equation 2 below:

$$euc= \sqrt{(a_1-b_1)^2+\cdots+(a_n-b_n)^2} \qquad (2)$$

Where a = a1,a2, ..., an, and b = b1, b2, ..., bn represents n attribute values of the two records, for attributes with category values.

Implementation of the K- Nearest Neighbor algorithm to the collected dataset was in the form of comments from students about online lectures. In data training, the part of the dataset was trained to make predictions or perform functions from the K-Nearest Neighbor algorithm. Instructions were provided through algorithms so that the machines we train can find their own correlations or learn patterns from the data provided [30]. At the training stage, the process was carried out using Python. At the data testing stage, the dataset was tested to see its accuracy, or in other words to look at its performance, namely the accuracy of the K-Nearest Neighbor algorithm.

E. Implementation of Feature Selection

Feature selection is one of the focuses in data mining. Feature selection is a process of selecting part of the variables of all variables in the dataset [31]. One of the feature selection methods is forward selection (sasongko). The feature selection method is used to select the most relevant features. After obtaining the most relevant features or attributes based on feature selection, the features and records are then processed using the k-nearest neighbor machine learning method to bring up the value of the performance of the model being tested [8]. The study used three *feature selections*, namely *Forward Selection* [32], *Genetic Algorithm* [33], and *Adaboost* [34].

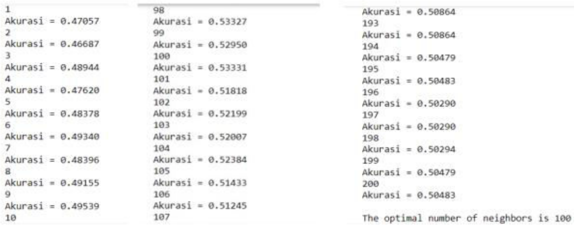
F. Evaluation

The evaluation is carried out to show the changes that have occurred in the model classically, in this case K-NN, with the model that has been optimized using *feature selection*.

RESULTS AND DISCUSSION

A. Implementation of the K-Nearest Neighbor Algorithm

The number of training data used in this study was 70% of the dataset. At this training stage, the process was carried out by using Python. Figure 9 is the result of the K-NN process using Python.



```
1
Akurasi = 0.47057
2
Akurasi = 0.46687
3
Akurasi = 0.48944
4
Akurasi = 0.47620
5
Akurasi = 0.48378
6
Akurasi = 0.49340
7
Akurasi = 0.48396
8
Akurasi = 0.49155
9
Akurasi = 0.49539
10
```

```
98
Akurasi = 0.53327
99
Akurasi = 0.52950
100
Akurasi = 0.53331
101
Akurasi = 0.51818
102
Akurasi = 0.52199
103
Akurasi = 0.52007
104
Akurasi = 0.52384
105
Akurasi = 0.51433
106
Akurasi = 0.51245
107
```

```
Akurasi = 0.50864
193
Akurasi = 0.50864
194
Akurasi = 0.50479
195
Akurasi = 0.50483
196
Akurasi = 0.50290
197
Akurasi = 0.50290
198
Akurasi = 0.50294
199
Akurasi = 0.50479
200
Akurasi = 0.50483

The optimal number of neighbors is 100
```

Figure 10. K-NN accuracy results

Testing of training data was carried out 200 times. The test results obtained the highest accuracy value in the 100th test, with a value of 53.33%.

```
Out[13]: Text(0, 0.5, 'Akurasi')
```

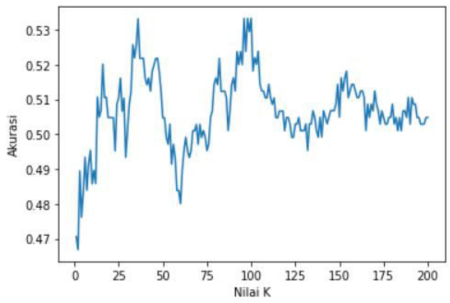
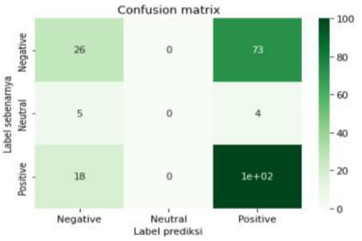
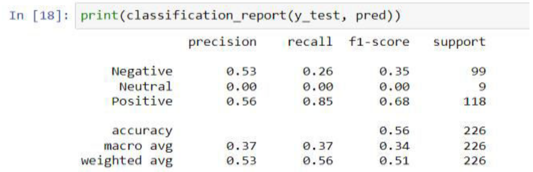


Figure 11. K-NN Accuracy Chart

A graph of its accuracy can be seen in figure 11. above, where it can be seen that the highest accuracy is found in the 100th test and the value is 0.53.

At the data testing stage, the dataset was tested to see its accuracy, or in other words, to see its performance, namely the accuracy of the K-Nearest Neighbor algorithm, and the combination of the K-NN Algorithm with feature selection, namely K-NN + Forward Selection, K-NN + Adaboost, and K-NN + Genetic Algorithm. Evaluation was carried out to show the changes that have occurred in the model classically, in this case K-NN, with the model that had been optimized using feature selection.

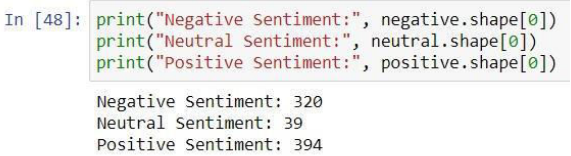




```
In [18]: print(classification_report(y_test, pred))

              precision    recall  f1-score   support

    Negative       0.53      0.26      0.35        99
     Neutral       0.00      0.00      0.00         9
    Positive       0.56      0.85      0.68       118

    accuracy                           0.56       226
   macro avg       0.37      0.37      0.34       226
weighted avg       0.53      0.56      0.51       226
```

Figure 12 Measurement of the accuracy or accuracy of prediction results

Figure 12 is the result of measuring the accuracy and accuracy of the prediction results. It is found that the accuracy is 56%.



```
In [48]: print("Negative Sentiment:", negative.shape[0])
         print("Neutral Sentiment:", neutral.shape[0])
         print("Positive Sentiment:", positive.shape[0])

         Negative Sentiment: 320
         Neutral Sentiment: 39
         Positive Sentiment: 394
```

Figure 13 Classification of Online Lecture Sentiments

Figure 13 is the result of the classification of student comments with online lecture topics, namely 394 positive sentiments, 320 negative, and 39 neutral comments.

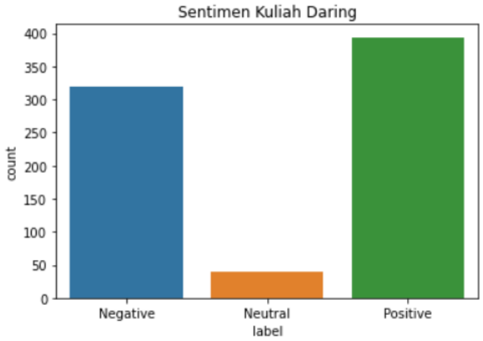


Figure 14. Online Lecture Sentiment Histogram

Figure 14 is a visual data in the form of a histogram that shows the results of the classification of student comments on the topic of online lectures. It can be seen that the number of negative comments is almost close to the number of positive comments, while the number of neutral comments is small.

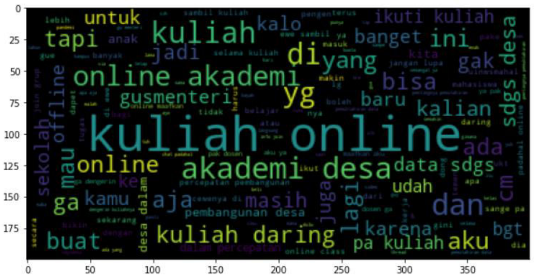


Figure 15. Positive Sentiment *of Word Cloud* Online Lecture

Figure 15 is a visualization in the form of a word cloud which shows the words often appearing in online lecture topics on positive sentiments.

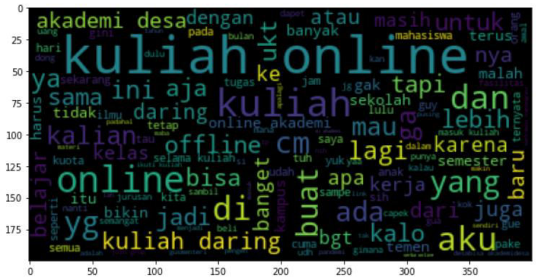


Figure 16. Negative Sentiment of *Word Cloud* Online Lecture

Figure 16 is a visual data in the form of word cloud which indicates the words that often emerge in online lecture topics on negative sentiments.

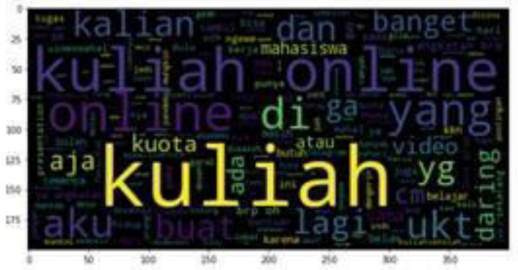
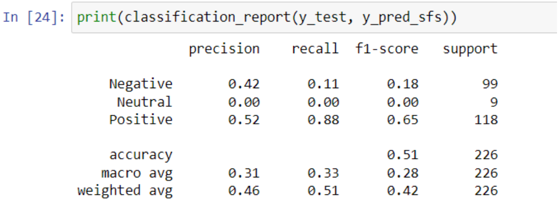


Figure 17. Neutral Sentiment of *Word Cloud* Online Lecture

Figure 17 is a visualization in the form of a word cloud that presents the words that often appear in topics of online lecture on neutral sentiments.

B. Feature Selection

At the phase of data testing, the dataset was tested to see its accuracy, or in other words, to find out its performance, namely the accuracy of the K-Nearest Neighbor algorithm, and the combination of the K-NN Algorithm with feature selection, namely K-NN + Forward Selection, K-NN + Adaboost, and K-NN + Genetic Algorithm. Evaluation was performed to show the changes that happened in the model classically, in this case K-NN, with the model that had been optimized using feature selection.



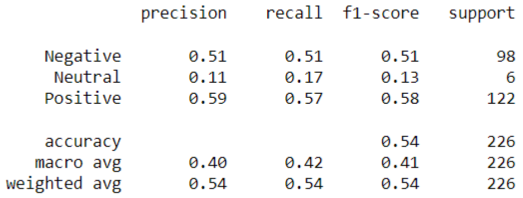
```
In [24]: print(classification_report(y_test, y_pred_sfs))

              precision    recall  f1-score   support

    Negative       0.42      0.11      0.18        99
     Neutral       0.00      0.00      0.00         9
    Positive       0.52      0.88      0.65       118

    accuracy                           0.51       226
   macro avg       0.31      0.33      0.28       226
weighted avg       0.46      0.51      0.42       226
```

Figure 18. *Feature Selection using Forward Selection*

The results of *feature selection* using *forward selection* obtained an accuracy result of 51%.

Figure 19. *Feature Selection* using Adaboost

The second feature selection result by Adabost was 54%.



```
              precision    recall  f1-score   support

    Negative       0.51      0.51      0.51        98
     Neutral       0.11      0.17      0.13         6
    Positive       0.59      0.57      0.58       122

    accuracy                           0.54       226
   macro avg       0.40      0.42      0.41       226
weighted avg       0.54      0.54      0.54       226
```

Figure 20. *Feature Selection using Genetic Algorithms*

The result of the third *feature selection* with a genetic algorithm was 55%.

C. Evaluation

The results of calculating the accuracy of the K-NN Algorithm with 3 *feature selection*s can be seen in the following table:

Table 1. Comparison results of k-nn algorithm with *feature selection*

Algorithm	Precision	Recall	Accuracy
KNN	0,37	0,37	56%
KNN + *Forward Selection*	0,31	0,33	51%
KNN + *Adaboost*	0,40	0,42	54%
KNN + *Genetic Algorithm*	0,36	0,32	55%

From Table 1 it can be seen that the precision and recall results of the K-NN Algorithm are 37% while the accuracy is 56%. The result of the K-NN Algorithm + *Forward Selection* Algorithm precision is 31%, followed by recall which is 33% and the accuracy itself is 51%. The precision result of the K-NN + Adaboost Algorithm is 40%, recall is 42% and accuracy is 54%. The precision result of the K-NN Algorithm + *Genetic Algorithm* is 36%, recall is 22% and accuracy is 55%. The initial hypothesis was that the use of feature selection is expected to increase the accuracy of the K-NN algorithm, but after the combination of the K-NN Algorithm with feature selection, the result is that the accuracy decreases. The possible cause is that the feature selection used does not match the K-NN algorithm.

From a total of 780 twitter data, those passing the preprocessing were 753 comments. Of the 753 comments, it showed that the percentage of student sentiment for the class was positive sentiment of 52% or 394 tweets, meanwhile the student sentiment for the class was negative sentiment of 42% or 320 tweets, and the rest of the student sentiment for the neutral sentiment class was 5% or 39 tweets. Therefore, the results of student opinions are on POSITIVE sentiments.

Factors that affect the level of accuracy include the amount of data and the number of k. research conducted by [35] used data of 1825 tweets and the optimal k value was 10 out of 20. The accuracy obtained was 84.65%, the precision was 87% and the recall was 86%, the f-measure was 87%. Another study conducted by [36] using data from 1000 tweets with optimal k is 1, resulting in an accuracy of 94.50%. In the research conducted using data from 780 tweets, the results obtained were quite far from previous studies, namely 56%. This is because the data used is less than previous research.

CONCLUSION

Based on the results of online lecture sentiment analysis research using the K-Nearest Neighbour Algorithm Based on Feature Selection that had been carried out to753 comments out of 780 tweets data, a classification of 394 positive sentiments, 320 negative sentiments, and 39 neutral sentiments was obtained. Therefore, the results of student opinions are on POSITIVE sentiments. Then the K-NN Algorithm test of training data was carried out 200 times. The test results obtained the highest accuracy value in the 100th test, namely with a value of 53.33%. Furthermore, the results of the precision and recall of the K-NN Algorithm are 37%, and the accuracy is 56%. The precision result of the K-NN + Forward Selection Algorithm is 31%, recall is 33% and accuracy is 51%. The precision result of the K-NN + Adaboost Algorithm is 40%, recall is 42% and accuracy is 54%. The precision result of the K-NN + Genetic Algorithm Algorithm is 36%, recall is 22% and accuracy is 55%.

ACKNOWLEDGMENTS

Acknowledgments are given to the Yayasan Komputasi Riau which has funded this research and thank to Lembaga Penelitian dan Pengabdian Masyarakat STMIK Amik Riau which has facilitated so that this research activity can be carried out (02/Kontrak/LPPM-STMIK Amik Riau/VIII/2021).

REFERENCES

[1] O. I. Handarini and S. S. Wulandari, “Pembelajaran Daring Sebagai Upaya Study From Home (SFH) Selama Pandemi Covid 19,” *Jurnal Pendidikan Administrasi Perkantoran (JPAP)*, vol. 8, no. 3, pp. 496–503, 2020, doi: 10.26740/jpap.v8n3.p496-503.

[2] H. A. Maulana and M. Hamidi, “Persepsi Mahasiswa terhadap Pembelajaran Daring pada Mata Kuliah Praktik di Pendidikan Vokasi,” *Equilibrium: Jurnal Pendidikan*, vol. 8, no. 2, pp. 224–231, 2020, doi: 10.26618/equilibrium.v8i2.3443.

[3] S. R. I. Rezeki, “Penggunaan Sosial Media Twitter dalam Komunikasi Organisasi (Studi Kasus Pemerintah Provinsi Dki Jakarta Dalam Penanganan Covid-19),” *Journal of Islamic and Law Studies*, vol. 04, no. 02, pp. 63–78, 2020.

[4] K. Curran, K. O’Hara, and S. O’Brien, “The role of twitter in the world of business,” *International Journal of Business Data Communications and Networking*, vol. 7, no. 3, pp. 1–15, 2011, doi: 10.4018/jbdcn.2011070101.

[5] S. Sendari, I. A. E. Zaeni, D. C. Lestari, and H. P. Hariyadi, “Opinion Analysis for Emotional Classification on Emoji Tweets using the Naïve Bayes Algorithm,” *Knowledge Engineering and Data Science*, vol. 3, no. 1, pp. 50–59, 2020, doi: 10.17977/um018v3i12020p50-59.

[6] M. K. Anam, “Analisis Respons Netizen Terhadap Berita Politik Di Media Online,” *Jurnal Ilmiah Ilmu Komputer*, vol. 3, no. 1, pp. 14–21, 2017, doi: 10.35329/jiik.v3i1.62.

[7] M. K. Anam, M. I. Mahendra, W. Agustin, Rahmaddeni, and Nurjayadi, “Framework for Analyzing Netizen Opinions on BPJS Using Sentiment Analysis and Social Network Analysis (SNA),” *Intensif*, vol. 6, no. 1, pp. 2549–6824, 2022, doi: 10.29407/intensif.v6i1.15870.

[8] M. K. Anam, B. Nanti, P. Gulo, M. B. Firdaus, and S. Erlinda, “Penerapan Naïve Bayes Classifier , K-Nearest Neighbor dan Decision Tree untuk Menganalisis Sentimen pada Interaksi Netizen dan Pemeritah Applications of Naïve Bayes Classifier , K-Nearest Neighbor and Decision Tree to Analyze Sentiment on Netizen and Gove,” *Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, vol. 21, no. 1, pp. 139–150, 2021, doi: 10.30812/matrik.v21i1.1092.

[9] A. P. Natasuwarna, “Analisis Sentimen Keputusan Pemindahan Ibukota Negara Menggunakan Klasifikasi Naive Bayes,” *Seminar Nasional Sistem Informasi dan Teknik Informatika*, pp. 47–53, 2019.

[10] D. Cahyanti, A. Rahmayani, and S. A. Husniar, “Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara,” *Indonesian Journal of Data and Science*, vol. 1, no. 2, pp. 39–43, 2020, doi: 10.33096/ijodas.v1i2.13.

[11] R. Sanjaya and F. Fitriyani, “Prediksi Bedah Toraks Menggunakan Seleksi Fitur Forward Selection dan K-Nearest Neighbor,” *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, vol. 5, no. 3, p. 316, 2019, doi: 10.26418/jp.v5i3.35324.

[12] R. Yunus, U. Ulfa, and M. D. Safitri, “Application of the K-Nearest Neighbors (K-NN) Algorithm for Classification of Heart Failure,” *Journal of Applied Intelligent System*, vol. 6, no. 1, pp. 1–9, 2021, doi: 10.33633/jais.v6i1.4513.

[13] Tanti and P. Sirait, “Optimalisasi Kinerja Klasifikasi Melalui Seleksi Fitur dan AdaBoost dalam Penanganan Ketidakseimbangan Kelas,” vol. 5, pp. 1377–1385, 2021, doi: 10.30865/mib.v5i4.3280.

[14] S. Mulyati, Y. Yulianti, and A. Saifudin, “Penerapan Resampling dan Adaboost untuk Penanganan Masalah Ketidakseimbangan Kelas Berbasis Naïve Bayes pada Prediksi Churn Pelanggan,” *Jurnal Informatika Universitas Pamulang*, vol. 2, no. 4, p. 190, 2017, doi: 10.32493/informatika.v2i4.1440.

[15] M. R. Fanani, “Penggabungan Forward Selection untuk Pemilihan Fitur pada Prediksi Bimbingan Konseling Siswa dengan Menggunakan Algoritma Naive Bayes,” *Smart Comp :Jurnalnya Orang Pintar Komputer*, vol. 9, no. 2, pp. 85–88, 2020, doi: 10.30591/smartcomp.v9i2.1924.

[16] S. F. Pane, R. Maulana Awangga, E. V. Rahcmadani, and S. Permana, “Implementasi Algoritma Genetika Untuk Optimalisasi Pelayanan Kependudukan,” *Jurnal Tekno Insentif*, vol. 13, no. 2, pp. 36–43, 2019, doi: 10.36787/jti.v13i2.130.

[17] E. S. Wahyuni, “Penerapan Metode Seleksi Fitur Untuk Meningkatkan Hasil Diagnosis Kanker Payudara,” *Simetris : Jurnal Teknik Mesin, Elektro dan Ilmu Komputer*, vol. 7, no. 1, p. 283, 2016, doi: 10.24176/simet.v7i1.516.

[18] I. Santoso, Windu Gata, and Atik Budi Paryanti, "Penggunaan Feature Selection di Algoritma Support Vector Machine untuk Sentimen Analisis Komisi Pemilihan Umum," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 3, no. 3, pp. 364–370, 2019, doi: 10.29207/resti.v3i3.1084.

[19] M. Dennis, R. Rahmaddeni, F. Zoromi, and M. K. Anam, "Penerapan Algoritma Naïve Bayes Untuk Pengelompokkan Predikat Peserta Uji Kemahiran Berbahasa Indonesia," *Jurnal Media Informatika Budidarma*, vol. 6, no. 2, pp. 1183–1190, Apr. 2022, doi: 10.30865/mib.v6i2.3956.

[20] B. N. Pikir, M. K. Anam, H. Asnal, Rahmaddeni, and T. A. Fitri, "Sentiment Analysis of Technology Utilization by Pekanbaru City Government Based on Community Interaction in Social Media," *JAIA – Journal Of Artificial Intelligence And Applications*, vol. 2, no. 1, pp. 32–40, 2021.

[21] A. F. Hidayatullah and M. R. Ma'arif, "Pre-processing Tasks in Indonesian Twitter Messages," in *Journal of Physics: Conference Series*, 2017. doi: 10.1088/1742-6596/755/1/011001.

[22] E. S. Romaito, M. K. Anam, Rahmaddeni, and A. N. Ulfah, "Perbandingan Algoritma SVM Dan NBC Dalam Analisa Sentimen Pilkada Pada Twitter," *CSRID Journal*, vol. 13, no. 3, pp. 169–179, 2021, doi: 10.22303/csrid.13.3.2021.169-179.

[23] M. K. Anam, Rahmaddeni, M. B. Firdaus, H. Asnal, and Hamdani, "Sentiment Analysis to analyze Vaccine Enthusiasm in Indonesia on Twitter Social Media," *JAIA – Journal Of Artificial Intelligence And Applications*, vol. 1, no. 2, pp. 23–27, 2021.

[24] P. WiraBuana, S. Jannet D.R.M., and I. Ketut Gede Darma Putra, "Combination of K-Nearest Neighbor and K-Means based on Term Re-weighting for Classify Indonesian News," *Int J Comput Appl*, vol. 50, no. 11, pp. 37–42, 2012, doi: 10.5120/7817-1105.

[25] R. Rahmiati, D. Irfan, A. Agustin, and S. Hediyati, "Aplikasi Pengukur Tingkat Sentimen Pelanggan Berdasarkan Komplain Pelanggan Pln Menggunakan Algoritma K-Nearest Neighbor," *INOVTEK Polbeng - Seri Informatika*, vol. 5, no. 2, p. 332, 2020, doi: 10.35314/isi.v5i2.1467.

[26] A. N. Ulfah and M. K. Anam, "Analisis Sentimen Hate Speech Pada Portal Berita Online Menggunakan Support Vector Machine (SVM)," *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, vol. 7, no. 1, pp. 1–10, 2020, doi: 10.35957/jatisi.v7i1.196.

[27] A. A. Maarif, "Penerapan Algoritma TF-IDF untuk Pencarian Karya Ilmiah," *Dokumen Karya Ilmiah | Tugas Akhir | Program Studi Teknik Informatika - S1 | Fakultas Ilmu Komputer | Universitas Dian Nuswantoro Semarang*, no. 5, p. 4, 2015.

[28] A. Pamuji, "Performance of the K-Nearest Neighbors Method on Analysis of Social Media Sentiment," *Juisi*, vol. 07, no. 01, pp. 32–37, 2021.

[29] Okfalisa, I. Gazalba, Mustakim, and N. G. I. Reza, "Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification," in *Proceedings - 2017 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2017*, 2018, vol. 2018-Janua, pp. 294–298. doi: 10.1109/ICITISEE.2017.8285514.

[30] S. Kumar and I. Chong, "Correlation analysis to identify the effective data in machine learning: Prediction of depressive disorder and emotion states," *Int J Environ Res Public Health*, vol. 15, no. 12, 2018, doi: 10.3390/ijerph15122907.

[31] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers and Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014, doi: 10.1016/j.compeleceng.2013.11.024.

[32] A. Rakhman and M. Rifqi Tsani, "Analisis Sentimen Review Media Massa Menggunakan Metode C4.5 Berbasis Forward Selection," *Smart Comp*, vol. 8, no. 2, pp. 78–82, 2019, doi: 10.30591/smartcomp.v8i2.1491.

[33] R. Wati, "Penerapan Algoritma Genetika Untuk Seleksi Fitur Pada Analisis Sentimen Review Jasa Maskapai Penerbangan," *Jurnal Evolusi*, vol. 4, no. 1, pp. 25–31, 2016.

[34] A. Andreyestha and A. Subekti, "Analisa Sentiment Pada Ulasan Film Dengan Optimasi Ensemble Learning," *Jurnal Informatika*, vol. 7, no. 1, pp. 15–23, 2020, doi: 10.31311/ji.v7i1.6171.

[35] A. R. Isnain, J. Supriyanto, and M. P. Kharisma, "Implementation of K-Nearest

Neighbor (K-NN) Algorithm For Public Sentiment Analysis of Online Learning," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 15, no. 2, p. 121, Apr. 2021, doi: 10.22146/ijccs.65176.

[36] M. Furqan, S. Mayang Sari, and P. Ilmu Komputer Fakultas Sains dan Teknologi, "Analisis Sentimen Menggunakan K-Nearest Neighbor Terhadap New Normal Masa Covid-19 Di Indonesia Sentiment Analysis using K-Nearest Neighbor towards the New Normal During the Covid-19 Period in Indonesia," 2022. [Online]. Available: www.tripadvisor.com