ILKOM Jurnal Ilmiah Vol. 15, No. 2, August 2023, pp. 290-302
Accredited 2nd by RISTEKBRIN No. 200/M/KPT/2020; E-ISSN 2548-7779 | P-ISSN 2087-1716

Research Article **Open Access (CC–BY-SA)**

Sentiment Analysis for Online Learning using The Lexicon-Based Method and The Support Vector Machine Algorithm

M. Khairul Anam [a,1,*]; Triyani Arita Fitri [a,2]; Agustin [a,3]; Lusiana [a,4]; Muhammad Bambang Firdaus [b,5]; Agus Tri Nurhuda [a,6]
[a] *STMIK Amik Riau, Jl. Purwodadi Indah Km. 10 Panam, Pekanbaru 28294, Indonesia*
[b] *University of Mulawarman, Jl. Sambaliung, Sempaja Sel, Kota Samarinda, 75242, Indonesia*
[1]*khairulanam@sar.ac.id;* [2]*Triyani@stmik-amik-riau.ac.id;* [3]*agustin@sar.ac.id;* [4]*lusiana@sar.ac.id;* [5]*bambangf@fkti.unmul.ac.id;* [6]*agustrinurhuda26@gmail.com*
** Corresponding author*

Article history: Received January 16, 2023; Revised March 20, 2023; Accepted July 17, 2023; Available online August 16, 2023

Abstract

The pros and cons regarding online learning has been a hot topic in society, both on social media and in the real world. Indonesian netizens still post opinions about online learning on social media such as Twitter. This study aims to analyze public comments to determine whether the trend of the comments is positive, negative, or neutral. The classification of netizen opinions is called sentiment analysis. This study applies 2 ways of carrying out sentiment analysis. The first stage employs the SVM algorithm with data labeling automatically obtained from the Emprit Academy drone portal while the second stage is still using the SVM algorithm but the data labeling with lexicon-based method. The results of this study are comparisons of labels obtained automatically from the drone Emprit Academy portal and labeling using lexicon based. The SVM algorithm obtains an accuracy of 90%, while the use of lexicon-based increases the accuracy value by 5% to 95%. It can be concluded that labeling data using a lexicon-based method can improve the accuracy of the SVM algorithm.

Keywords: Lexicon Based; Online Learning; Sentiment Analysis; SVM.

Introduction

The online learning process is one way to overcome the problems of the 2019 COVID pandemic and distance learning [1]. However, online learning raises problems such as the absence of devices owned by students to implement this learning system [2]. The pros and cons related to online learning have been widely raised, both in media such as television and social media such as Twitter. Twitter is a social networking service that allows users to send text, images and videos called tweets [3]. Many researchers have studied tweets written on Twitter using both sentiment analysis [4] and Social Network Analysis (SNA) [5].

This study applies sentiment analysis to examine tweets. Sentiment analysis is a type of natural language processing [6]. This analysis examines the context to identify information from data sources [7]. By employing, sentiment analysis, artificial intelligence can understand the emotions of a brand, product or even a certain social situation [8], [9]. Previous research conducted sentiment analysis using several algorithms such as Naïve Bayes [10], Adaboost [11], Support Vector Machine (SVM) [12], KNN [13], Decision Tree [14], etc.

This research applies the SVM algorithm. In classification modeling, SVM has a more mature and clearer concept mathematically compared to other classification techniques [15]. This technique can also solve classification and regression problems with linear and nonlinear [16]. In addition, SVM has a high level of accuracy. Research conducted by [17] in a classification experiment using the SVM algorithm with the Radial Basis Function (RBF) kernel obtained an accuracy of 88%. Another research that analyzed reviews on restaurants using the SVM technique obtained the highest accuracy of 79%. Furthermore, [18] conducted a comparison of the kernels on SVM, namely RBF, Linear, and Multinomial and produced that SVM with the Linear kernel obtained the highest accuracy of 85.6%.

http://dx.doi.org/10.33096/ilkom.v15i2.1590.290-302

Based on some of these researches, this study will use the SVM algorithm with the linear kernel. This study applies 2 classification methods, namely data classification on the Drone Emprit Academy (DEA) portal and the lexicon-based method. The lexicon-based method is an approach in sentiment analysis that uses a dictionary or lexicon that contains a list of words with associated sentiment values [19]. This method aims to identify and assess the sentiments in the text by matching the words found in the text with the sentiment dictionary [20]. The result of this study is a comparison of the classifications produced by the DEA portal and the lexicon-based method.

Methods

Research methodology is a sequence in conducting research that is intended to prepare research more conceptual and well directed according to the research objectives in order to produce a system that has been tested and in turn may solve the problems. The flow or steps to be applied in this study can be seen in the following **Figure 1**.

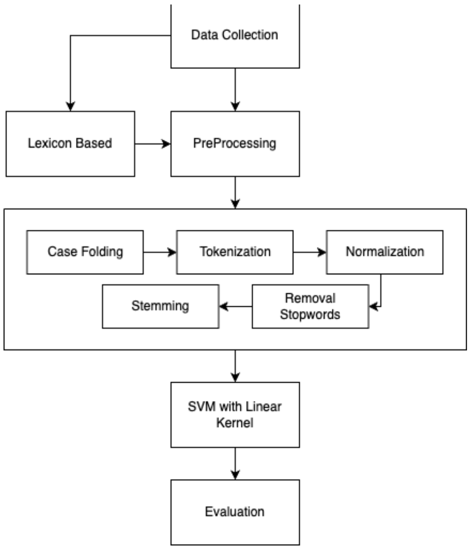

Figure 1. Research Stages

A. *Data Collection*

The data collection process was carried out by accessing the DEA portal of Universitas Islam Indonesia, a system capable of monitoring and analyzing social media on various online platforms based on Artificial Intelligence (AI) technology and Natural learning process (NLP). The data obtained was 1225 tweets in Indonesian language which were then divided into training and testing data with a ratio of 80%:20% respectively. All Twitter data is downloaded and stored in the .xlxs document for further processing. **Figure 2** is Twitter data stored in .csv format.

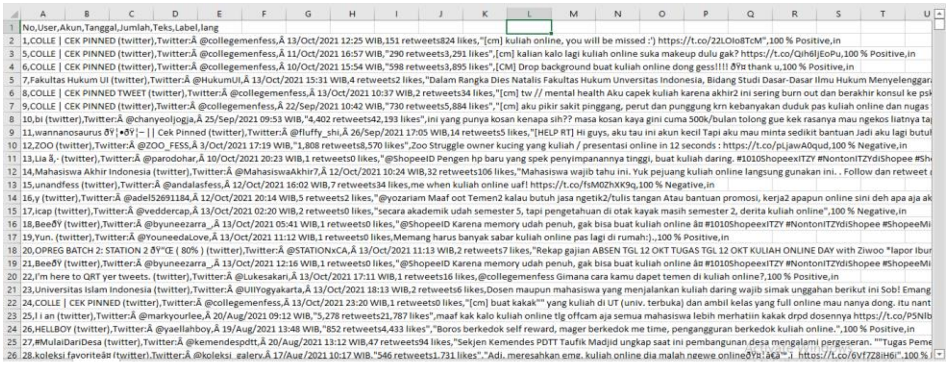

Figure 2. Tweets data in.csv format

This data has not gone through the pre-processing process so it is still mixed with other characters attached to the data. The type of data in this study is quantitative data obtained within 13 months from March 2021 to April 2022. The proportion of data for data testing and training is 20% and 80% respectively. The training data will be used to train the algorithm in determining the appropriate model.

B. *Data Labelling Applying Lexicon-based Method*

Data labelling stages play an important role in the classification process because this research uses an approach at the word level where the data processed is the word to obtain sentiment scores. The lexicon dictionary used in this study is the InSet lexicon dictionary [21] which contains a list of words that contain both positive and negative sentiment and already has a weight of value for each word. The dictionary consists of 3609 positive and 6609 negative words. This research adds a few words related to the topic of Online College.

1	[bayang, doa, lanjut, kuliah, online, ajar, on...	-20	negatif
2	[anak, ringan, finansial, anak, malu, bayar, k...	-10	negatif
3	[praktikum, offline, kuliah, online, reguler, ...	0	netral
4	[weh, beneran, nih, full, kuliah, online, daft...	-1	negatif
...	...	...	...
1195	[maap, juragan, butuh, jasa, ketiktulis, tanga...	-14	negatif

Figure 3. Display of *inset Lexicon Dictionary*

Figure 3 shows the determination of labels for each word; positive, neutral or negative words. The next stage was the process of weighing each word in the preprocessed data using the InSet lexicon dictionary. **Figure 4** portraits the results of the adjustment process in the InSet lexicon dictionary with preprocessing data.

suka	drop	thank	hukum	studi	ilmu	selenggara	umum	sistem	tau	...	buram	antusias	hadiah	tunai	nila	inovasi	panjang	sindir	muslim	sentiment
0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	3
0	1	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
0	0	0	3	1	1	1	1	1	0	...	0	0	0	0	0	0	0	0	0	17
0	0	0	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0	0	0	2
0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	20
0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0	0	0	5
0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	-3
0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

Figure 4. Weighing Result

The weighting results are used to calculate the number of positive, neutral, and negative words. If the number of positive words is greater than the number of negative words, the sentiment label is positive (score 1); if the number of positive words is less than the number of negative words, the sentiment label is negative (score -1); and if the number of positive words equal to the number of negative words, the sentiment label is neutral (score 0). After carrying out the weighting, the next step is to conduct the lexicon-based addition process. **Figure 5** is the result of the classification process.

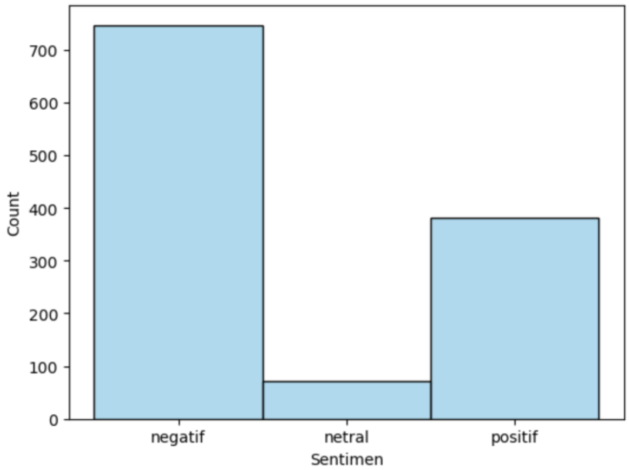

Figure 5. Result of Lexicon-Based Classification

Figure 5 shows the results of testing the Python programming using training data. These results indicate that the highest public opinion data is for the negative sentiment class, followed by the positive sentiment class, and finally the neutral sentiment class.

C. Preprocessing

There are several steps to be carried out in the Preprocessing stage, namely case folding and cleansing, tokenization, normalization, removal of stop words, and stemming. The purpose of the preprocessing stage is to eliminate problems that can interfere with the results of data processing.

1. *Case folding dan Cleaning*

Furthermore, the test data and training data were processed in the case folding stage where the data in the form of letter characters will be converted to lowercase. The data from the document cleansing process that has been acquired is cleaned from characters such as html, hashtags, username (@username), punctuation marks such as (. , ? ! [ ] / % : ; < > ( ) *)), numbers (0, 1,2,3,4,5,6,7,8,9,0), and other characters besides the alphabet. The purpose of the cleaning process is to reduce noise. This stage is very important to obtain valid data to be processed at a later stage [23]. **Table 1** displays the results of tweet comment data with the keywords of *perkuliahan daring*, online lectures, and *kuliah daring* that have undergone case folding and cleaning processes.

Table 1. Results of Case folding and Cleaning Process

No.	Before Case Folding and Cleaning Process	After Case Folding and Cleaning Process
1	*beneran, saya kuliah di UT, IP saya lumayan bagus Alhamdulillah, dan ya ngomongnya begini, bodoh banget, makin saya pamer IP, bikin mata pedih terus, Diperkirakan kuliah penuh* online *tanpa stres*	*beneran saya kuliah di ut ip saya lumayan bagus alhamdulillah dan ya ngomongnya begini bodoh banget makin saya pamer ip bikin mata pedih terus, diperkirakan kuliah penuh* online *tanpa stres*
2	*Kenapa hanya saat kuliah* online *saja mata kuliah tertentu bermasalah (sinyalnya laptop), dan hanya mata kuliah itu saja, yang lain tidak?*	*kenapa hanya saat kuliah online saja mata kuliah tertentu bermasalah sinyalnya laptop dan hanya mata kuliah itu saja yang lain tidak*
3	*Bersyukur sekali bukan, waktu awal masuk kuliah online banyak kendala dan yang pertama kuliah b.indo lalu ada kendala kebetulan belum ada lesnya jadi saya berinisiatif untuk ngobrol sama dosennya justru status sebagai mahasiswa baru membuatnya tertekan untuk mengatasinya*	*bersyukur sekali bukan waktu awal masuk kuliah online banyak kendala dan yang pertama kuliah bindo lalu ada kendala kebetulan belum ada lesnya jadi saya berinisiatif untuk ngobrol sama dosennya justru status sebagai mahasiswa baru membuatnya tertekan untuk mengatasinya*
4	*Sore ini saya pergi ke apotik untuk membeli obat batuk dan pilek yang tidak terkontrol, disitu saya bertemu ibu-ibu yang baru datang dan minta cepat karena katanya anaknya mau belajar* online *tapi batuk dan pileknya belum sembuh, keadaannya ada aku yang duluan.*	*sore ini saya pergi ke apotik untuk membeli obat batuk dan pilek yang tidak terkontrol disitu saya bertemu ibu-ibu yang baru datang dan minta cepat karena katanya anaknya mau belajar* online *tapi batuk dan pileknya belum sembuh keadaannya ada aku yang duluan*
5	*Setelah sekian bulan purnama, kita ketemu* offline *lagi dengan anak kampus yang belajar* online https://t.co/khk8xG68IZ	*setelah sekian bulan purnama kita ketemu* offline *lagi dengan anak kampus yang belajar* online

2. *Tokenization*

The tokenization process is the stage where the data which was originally in the form of a description is broken down into words. The tokenization process is to separate words from a sentence comment by marking (,) as a separator for each word [24]. The results of the tweet comment data that have gone through the case folding and cleansing process will be tokenized. Data resulted from Tokenization process is presented in **Table 2**.

Table 2. Results of Tokenization Process

No.	Before Tokenization Process	After Tokenization Process
1	*beneran saya kuliah di ut ip saya lumayan bagus alhamdulillah dan ya ngomongnya begini bodoh banget makin saya pamer ip bikin mata pedih terus, diperkirakan kuliah penuh* online *tanpa stres*	*beneran saya kuliah di ut ip saya lumayan bagus alhamdulillah dan ya ngomongnya begini bodoh banget makin saya pamer ip bikin mata pedih terus, diperkirakan kuliah penuh* online *tanpa stres*
2	*kenapa hanya saat kuliah online saja mata kuliah tertentu bermasalah sinyalnya laptop dan hanya mata kuliah itu saja yang lain tidak*	*kenapa hanya saat kuliah online saja mata kuliah tertentu bermasalah sinyalnya laptop dan hanya mata kuliah itu saja yang lain tidak*
3	*bersyukur sekali bukan waktu awal masuk kuliah online banyak kendala dan yang pertama kuliah bindo lalu ada kendala kebetulan belum ada lesnya jadi saya berinisiatif untuk ngobrol sama dosennya justru status sebagai mahasiswa baru membuatnya tertekan untuk mengatasinya*	*bersyukur sekali bukan waktu awal masuk kuliah online banyak kendala dan yang pertama kuliah bindo lalu ada kendala kebetulan belum ada lesnya jadi saya berinisiatif untuk ngobrol sama dosennya justru status sebagai mahasiswa baru membuatnya tertekan untuk mengatasinya*
4	*sore ini saya pergi ke apotik untuk membeli obat batuk dan pilek yang tidak terkontrol disitu saya bertemu ibu-ibu yang baru datang dan minta cepat karena katanya anaknya mau belajar* online *tapi batuk dan pileknya belum sembuh keadaannya ada aku yang duluan*	*sore ini saya pergi ke apotik untuk membeli obat batuk dan pilek yang tidak terkontrol disitu saya bertemu ibu-ibu yang baru datang dan minta cepat karena katanya anaknya mau belajar* online *tapi batuk dan pileknya belum sembuh keadaannya ada saya yang duluan*

No.	Before Tokenization Process	After Tokenization Process
5	*setelah sekian bulan purnama kita ketemu* offline *lagi dengan anak kampus yang belajar* online	*setelah sekian bulan purnama kita ketemu* offline *lagi dengan anak kampus yang belajar* online

3. Word Normalization

Lexical normalization is the process of transforming tokens into a canonical form consistent with the dictionary and grammar. These tokens include words that are misspelt or intentionally shortened (elisions) due to character limit in case of Twitter [25]. For example, the word "*bgs*" is transformed into "*bagus*". **Table 3** shows the normalized data.

Table 3. Results of Data Normalization

No.	Before Normalization	After Normalization
1	*beneran saya kuliah di ut ip saya lumayan bagus alhamdulillah dan ya ngomongnya begini bodoh banget makin saya pamer ip bikin mata pedih terus, diperkirakan kuliah penuh* online *tanpa stres*	*[beneran, saya, kuliah, di, ut, ip, saya, lumayan, bagus, alhamdulillah, dan, ya, ngomongnya, begini, bodoh, banget, makin, saya, pamer, ip, bikin, mata, pedih, terus, diperkirakan, kuliah, penuh,* online*, tanpa, stres]*
2	*kenapa hanya saat kuliah* online *saja mata kuliah tertentu bermasalah sinyalnya laptop dan hanya mata kuliah itu saja yang lain tidak*	*[kenapa, hanya, saat, kuliah, online, saja, mata, kuliah, tertentu, bermasalah, sinyalnya, laptop, dan, hanya, mata, kuliah, itu, saja, yang, lain, tidak]*
3	*bersyukur sekali bukan waktu awal masuk kuliah online banyak kendala dan yang pertama kuliah bindo lalu ada kendala kebetulan belum ada lesnya jadi saya berinisiatif untuk ngobrol sama dosennya justru status sebagai mahasiswa baru membuatnya tertekan untuk mengatasinya*	*[bersyukur, sekali, bukan, waktu, awal, masuk, kuliah, online, banyak, kendala, dan, yang, pertama, kuliah, bindo, lalu, ada, kendala, kebetulan, belum, ada, lesnya, jadi, saya, berinisiatif, untuk, ngobrol, sama, dosennya, justru, status, sebagai, mahasiswa, baru, membuatnya, tertekan, untuk, mengatasinya]*
4	*sore ini saya pergi ke apotik untuk membeli obat batuk dan pilek yang tidak terkontrol disitu saya bertemu ibu-ibu yang baru datang dan minta cepat karena katanya anaknya mau belajar* online *tapi batuk dan pileknya belum sembuh keadaannya ada saya yang duluan*	*[sore, ini, saya, pergi, ke, apotik, untuk, membeli, obat, batuk, dan, pilek, yang, tidak, terkontrol, disitu, saya, bertemu, ibu-ibu, yang, baru, datang, dan, minta, cepat, karena, katanya, anaknya, mau, belajar,* online*, tapi, batuk, dan, pileknya, belum, sembuh, keadaannya, ada, saya, yang, duluan]*
5	*setelah sekian bulan purnama kita ketemu* offline *lagi dengan anak kampus yang belajar* online	*[setelah, sekian, bulan, purnama, kita, ketemu,* offline*, lagi, dengan, anak, kampus, yang, belajar,* online]

4. Stop words removal

Stop words removal is a stage of removing meaningless common words [26]. The results of normalized tweet comment data will go through the stop words removal process. In this process, the common words that are discarded are those in accordance to the KBBI. **Table 4** shows comment data that has undergone the Stop words removal process.

Table 4. Results of Stopwords Removal Process

No.	Before stop words removal	After stop words removal
1	*[beneran, saya, kuliah, di, ut, ip, saya, lumayan, bagus, alhamdulillah, dan, ya, ngomongnya, begini, bodoh, banget, makin, saya, pamer, ip, bikin, mata, pedih, terus, diperkirakan, kuliah, penuh,* online*, tanpa, stres]*	*beneran kuliah ut ip lumayan bagus alhamdulillah ya ngomongnya bodoh banget pamer ip bikin mata pedih terus, kuliah penuh* online *stres*
2	*[kenapa, hanya, saat, kuliah,* online*, saja, mata, kuliah, tertentu, bermasalah, sinyalnya, laptop, dan, hanya, mata, kuliah, itu, saja, yang, lain, tidak]*	*kuliah* online *mata kuliah bermasalah sinyalnya laptop mata kuliah*
3	*[bersyukur, sekali, bukan, waktu, awal, masuk, kuliah, online, banyak, kendala, dan, yang, pertama, kuliah, bindo, lalu, ada, kendala, kebetulan, belum, ada, lesnya, jadi, saya, berinisiatif, untuk, ngobrol, sama, dosennya, justru, status, sebagai, mahasiswa, baru, membuatnya, tertekan, untuk, mengatasinya]*	*bersyukur masuk kuliah online kendala kuliah bindo kendala lesnya berinisiatif ngobrol dosennya status mahasiswa membuatnya tertekan mengatasinya*
4	*[sore, ini, saya, pergi, ke, apotik, untuk, membeli, obat, batuk, dan, pilek, yang, tidak, terkontrol, disitu, saya, bertemu, ibu-ibu, yang, baru, datang, dan, minta, cepat, karena, katanya, anaknya, mau, belajar,* online*, tapi, batuk, dan, pileknya, belum, sembuh, keadaannya, ada, saya, yang, duluan]*	*sore pergi apotik membeli obat batuk pilek terkontrol disitu bertemu ibu-ibu cepat anaknya belajar* online *batuk pileknya sembuh keadaannya duluan*
5	*[setelah, sekian, bulan, purnama, kita, ketemu,* offline*, lagi, dengan, anak, kampus, yang, belajar,* online*]*	*sekian purnama ketemu* offline *anak kampus belajar* online

5. Stemming

The stemming process is turning words into basic words by removing affixes such as prefixes, infixes, suffixes and conflixes (combinations of prefixes and suffixes) in derived words that are matched with KBBI [27]. This process is the final process of the preprocessing stage. At this stage, each word is converted into a basic word that is adjusted to KBBI. **Table 5** shows the results of the tweet comment data that have undergone the Stemming process. The stemming process in this study applied literature in natural language and text processing to help analyze Indonesian texts more effectively. Literature technique provides a variety of features including word trimming, lower case conversion, and etc. This algorithm is very useful in text classification, opinion mining, information extraction, and other tasks that involve processing Indonesian [28].

Table 5. Results of Stemming Process

No.	Before	After
1	*beneran kuliah ut ip lumayan bagus alhamdulillah ya ngomongnya bodoh banget pamer ip bikin mata pedih terus, kuliah penuh* online *stres*	*beneran kuliah ut ip lumayan bagus alhamdulillah ya ngomong bodoh banget pamer ip bikin mata pedih terus kuliah penuh* online *stres*
2	*kuliah* online *mata kuliah bermasalah sinyalnya laptop mata kuliah*	*kuliah* online *mata kuliah masalah sinyal laptop mata kuliah*
3	*bersyukur masuk kuliah* online *kendala kuliah bindo kendala lesnya berinisiatif ngobrol dosennya status mahasiswa membuatnya tertekan mengatasinya*	*syukur masuk kuliah* online *kendala kuliah bindo kendala les inisiatif ngobrol dosen status mahasiswa buat tekan atas*
4	*sore pergi apotik membeli obat batuk pilek terkontrol disitu bertemu ibu-ibu cepat anaknya belajar* online *batuk pileknya sembuh keadaannya duluan*	*sore pergi apotik beli obat batuk pilek kontrol situ temu ibu cepat anak ajar* online *batuk pilek sembuh ada duluan*
5	*sekian purnama ketemu* offline *anak kampus belajar* online	*sekian purnama ketemu* offline *anak kampus ajar* online

D. Calculating Term Frequency-Inserve Document Frequency

Term Frequency-Inserve Document Frequency (TF-IDF) is the number of selected word frequencies/number of words. An example of the TF calculation can be seen in **Figure 6**. This study employed the TF-IDF vectorizer from the sklearn library. Following is the implementation of the TF-IDF vectorizer. After implementing the TF-IDF Vectorizer, the frequency or weight of each tweet comment data will be generated. The following is the weight that has been generated through the TF-IDF process.

```
(0, 85)     0.21906915932898333 (8, 51)     0.20359775019352028
(0, 63)     0.08871387357024871 (8, 77)     0.20359775019352028
(0, 67)     0.21906915932898333 (8, 75)     0.20359775019352028
(0, 95)     0.21906915932898333 (8, 72)     0.20359775019352028
(0, 66)     0.21906915932898333 (8, 58)     0.40719550038704055
(0, 54)     0.18622879958332572 (8, 99)     0.20359775019352028
(0, 16)     0.21906915932898333 (8, 1)      0.20359775019352028
(0, 64)     0.21906915932898333 (8, 44)     0.40719550038704055
(0, 9)      0.21906915932898333 (8, 91)     0.20359775019352028
(0, 18)     0.21906915932898333 (8, 65)     0.20359775019352028
(0, 56)     0.21906915932898333 (8, 47)     0.20359775019352028
(0, 103)    0.21906915932898333 (8, 92)     0.1730766883505759
(0, 4)      0.21906915932898333 (8, 63)     0.08244859808281284
(0, 8)      0.21906915932898333 (8, 43)     0.09924555050442142
```

Figure 6. Data Generated from TF-IDF Process

Equation 1 is the formula to calculate the IDF:

$$idf = log\,\frac{1+n}{1+df} + 1 \qquad (1)$$

The calculation from the formula above is the log of one plus number of documents divide by one plus document frequency, and the result is plus one. Formula to calculate the IDF is Log (n/df), for the word of "*syukur*", the calculation is given as follows:

$LOG\left(\frac{5+1}{1+1}\right) + 1 = 1{,}477121255.$

Table 6 displays all the results of IDF calculation.

Table 6. Calculation of IDF

Word	IDF
	Log (n/df)
syukur	LOG((5 + 1) / (1 + 1)) + 1 = 1,477121255
masuk	LOG((5 + 1) / (1 + 1)) + 1 = 1,477121255
kuliah	LOG((5 + 1) / (3 + 1)) + 1 = 1,176091259
online	LOG((5 + 1) / (5 + 1)) + 1 = 1
kendala	LOG((5 + 1) / (1 + 1)) + 1 = 1,477121255
bindo	LOG((5 + 1) / (1 + 1)) + 1 = 1,477121255
les	LOG((5 + 1) / (1 + 1)) + 1 = 1,477121255
inisiatif	LOG((5 + 1) / (1 + 1)) + 1 = 1,477121255
ngobrol	LOG((5 + 1) / (1 + 1)) + 1 = 1,477121255
dosen	LOG((5 + 1) / (1 + 1)) + 1 = 1,477121255
status	LOG((5 + 1) / (1 + 1)) + 1 = 1,477121255
mahasiswa	LOG((5 + 1) / (1 + 1)) + 1 = 1,477121255
buat	LOG((5 + 1) / (1 + 1)) + 1 = 1,477121255
tekan	LOG((5 + 1) / (1 + 1)) + 1 = 1,477121255
atas	LOG((5 + 1) / (1 + 1)) + 1 = 1,477121255

The calculation of TF-IDF is a step of classification process applying SVM method. The number of frequencies for each word influences the next classification process. Formula to calculate TF-IDF is given as follow Equation 2.

$$TF - IDF = tf_{ij} \times lo\,g\left(\frac{D}{df_j}\right) \quad (2)$$

Calculation was performed by multiplying the TF matrix value by the log division result of D with the row value of DF. The following is an example of a TF-IDF calculation:

$1 \times 1{,}47712125 = 1{,}47712125$

The TF-IDF of K1 to K5 from the preprocessing results can be seen in the following **Table 7**.

Table 7. Calculation of TF-IDF

word	*TF-IDF*				
	K1	***K2***	***K3***	***K4***	***K5***
syukur	0	0	1,47712125	0	0
masuk	0	0	1,47712125	0	0
kuliah	$2 \times 1{,}176091259 = 2{,}352182518$	$3\,x\,1{,}176091259 = 3{,}528273777$	2,352182518	0	0
online	$1 \times 1 = 1$	1	1	1	1
kendala	0	0	$2\,x\,1{,}47712125 = 2{,}954242509$	0	0
bindo	0	0	1,47712125	0	0
les	0	0	1,47712125	0	0
inisiatif	0	0	1,47712125	0	0
ngobrol	0	0	1,47712125	0	0
dosen	0	0	1,47712125	0	0
status	0	0	1,47712125	0	0
mahasiswa	0	0	1,47712125	0	0
buat	0	0	1,47712125	0	0

word	*TF-IDF*				
	K1	***K2***	***K3***	***K4***	***K5***
tekan	0	0	1,47712125	0	0
atas	0	0	1,47712125	0	0

E. Testing and Evaluation

At this stage, testing of the SVM algorithm is carried out by determining its precision, recall and accuracy. Precision is used to calculate predicted class accuracy according to the actual class for accuracy results. Recall, on the other hand, is used to measure the sensitivity of the measurement to the dataset or the predictive ability of the system according to the level of truth for recalling relevant documents. Tests were carried out to show the results obtained in tabular form which showed data in the form of positive, neutral and negative. The following are provisions for accuracy using the confusion matrix.

Table 8. *Confusion matrix*

Actual Data	Predicted Data		
	Negative	***Neutral***	***Positive***
Negative	TNg	NgN	FN
Neutral	NNg	TN	NP
Positive	FP	PN	TP

Table 8 of the confusion matrix above presents the prediction results using a classification technique to calculate the predicted class accuracy according to the actual class for accuracy results. To measure precision, the following Equation 3 is used:

$$Precision = \frac{True\ Positif}{True\ Positif + False\ Positif} \tag{3}$$

The calculation of precision for each word class applies the following Equation 4, 5, 6:

$$Positive = \frac{TP}{TP + NP + FN} \tag{4}$$

$$Neutral = \frac{TN}{TN + PN + NgN} \tag{5}$$

$$Negative = \frac{TNg}{TNg + FP + NNg} \tag{6}$$

Recall is used to measure the sensitivity of the measurement to the dataset or the predictive ability of the system according to the level of truth to retrieve the relevant documents for recall measurement for each word class. The recall formula uses the following Equation 7, 8, 9:

$$Positive = \frac{TP}{TP + NP + FP} \tag{7}$$

$$Neutral = \frac{TN}{TN + NP + NNg} \tag{8}$$

$$Negative = \frac{TNg}{TNg + FN + NgN} \tag{9}$$

Results and Discussion

This section describes the results and discussion which aims to address all the questions of this research which is the classification of public opinion on online lectures using the SVM Algorithm on Twitter Media.

A. Testing and Evaluation

The first test is the classification of data obtained directly from the DEA portal. Prior to the evaluation process, the data used needs to be balanced because the data obtained from the DEA data still seems very unbalanced. **Figure 7** is the result of data classification from the DEA.

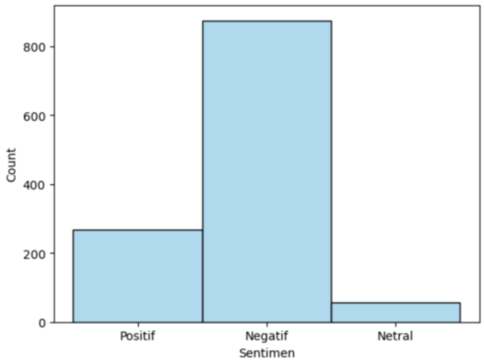

Figure 7. Classification of Data from DEA Portal

If these results are directly processed using the SVM algorithm, the accuracy value obtained would be less optimal. This study uses the Synthetic Minority Over-sampling Technique (SMOTE) method to balance the data. **Figure 8** is a balancing of classification data from DEA portal.

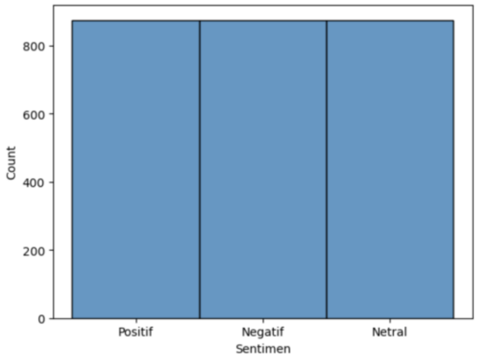

Figure 8. Balancing of Classification Data from DEA Portal

The accuracy results obtained after balancing the data are adequately high which is 90%. **Figure 9** shows the result of SVM accuracy.

```
Accuracy Score untuk Support Vector Machine Model ::  0.900952380952381
              precision    recall  f1-score   support

     Negatif       0.97      0.79      0.87       192
      Netral       0.97      0.95      0.96       165
     Positif       0.79      0.98      0.87       168

    accuracy                           0.90       525
   macro avg       0.91      0.91      0.90       525
weighted avg       0.91      0.90      0.90       525
```

Figure 9. Results of SVM Accuracy

Figure 9 shows that the values for precision, recall, and F1 of 91%, 91%, and 90% respectively. **Figure 10** is a confusion matrix of data classification from the DEA portal.

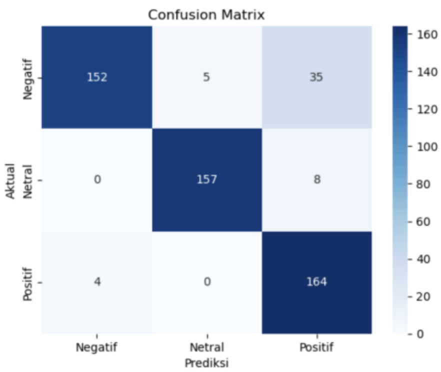

Figure 10. Confusion Matrix of SVM

The next step is to design SVM model using the lexicon-based method. **Figure 5** shows the results of the lexicon-based classification which indicates that the positive class has increased compared to the positive class classification obtained directly from the DEA portal. However, the data obtained is not yet balanced, so data balancing is still needed at this stage. **Figure 11** shows the results of balancing data using the SMOTE method.

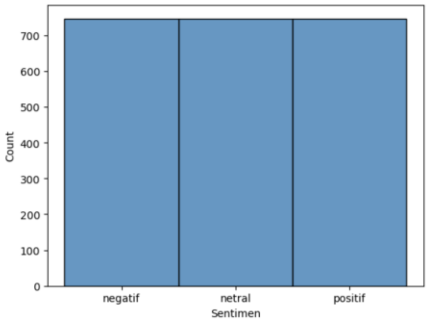

Figure 11. Results of Data Balancing using Lexicon-Based classification Method

After balancing the data using the SMOTE method, **Figure 12** presents the accuracy results obtained using the SVM algorithm.

```
Accuracy Score untuk Support Vector Machine Model ::  0.9465478841870824
              precision    recall  f1-score   support

     negatif       0.92      0.94      0.93       153
      netral       0.95      1.00      0.98       144
     positif       0.97      0.90      0.94       152

    accuracy                           0.95       449
   macro avg       0.95      0.95      0.95       449
weighted avg       0.95      0.95      0.95       449
```

Figure 12. The Accuracy Results of Lexicon-Based using SVM algorithm

The accuracy results increased from 90% to 95%. In addition, the value of precision, recall, and F1-Score increased to 95%. The results of the confusion matrix can be seen in **Figure 13**.

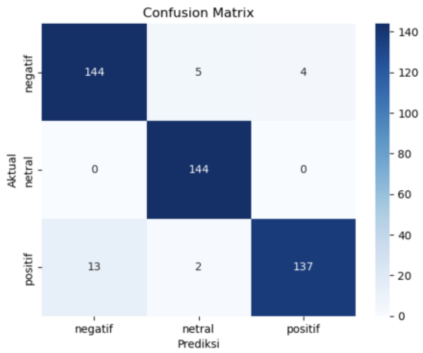

Figure 13. Confusion Matrix of SVM + Lexicon Based

B. Discussion

This research is carried out through several stages. The first stage was data collection on the DEA portal. The data obtained is automatically labeled positive, negative, and neutral. However, the data is not yet balanced, so a data balancing process is needed. To balance the data, this study employed the SMOTE method which is one of the oversampling techniques used in processing unbalanced data in machine learning. This method aims to overcome the class imbalance problem by making synthetic samples from the minority class [29]. The accuracy value before the data balancing process is 72%. These results were obtained using data with automatic labeling from the DEA. After the data balancing process, the accuracy results increased to 90%. This has also been experienced in research conducted by [30] who applied the SMOTE method on the SVM algorithm which was able to increase accuracy to 93%. However, there are also studies that used the SMOTE method which results in a decreased accuracy value. In a study conducted by [31], the accuracy value of applying the SVM SMOTE method reduced the accuracy value from 79% to 77%.

This study applied the Lexicon-Based method for re-labeling using the Indonesian dictionary which is applied to this method. This approach can also increase the value of data accuracy obtained from the DEA portal from 70% to 82%. The accuracy value increased to 95% after balancing the data.

Table 9. Comparison of Accuracy Score

Algorithm	Accuracy	Precision	Recall	F1-Score
SVM	72%	64%	72%	67%
SVM SMOTE	82%	83%	82%	79%
Lexicon Based SVM	90%	91%	91%	90%
Lexicon Based SVM SMOTE	95%	95%	95%	95%

Conclusion

The labeling process using lexicon-based can increase the accuracy value. Furthermore, the SMOTE method in this study can also improve accuracy. This study applied 80:20 data splitting for training and testing data respectively and employed the Linear kernel SVM. Therefore, future research is expected to apply various data splitting to obtain the best splitting. In addition, future research can also use other SVM kernels to determine the level of accuracy produced from applied SVM kernels techniques.

References

[1] Jamila, Ahdar, and E. Natsir, "Problematika Guru dan Siswa dalam Proses Pembelajaran Daring pada Masa Pandemi COVID-19 di UPTD SMP Negeri 1 Parepare," *AL MA'ARIEF: Jurnal Pendidikan Sosial dan Budaya*, vol. 3, no. 2, pp. 101–110, 2021, doi: 10.35905/almaarief.v3i2.2346.

[2] J. Suprapmanto and Utomo, "Analysis of Online Learning Problems During The COVID-19 Pandemic and Their Solutions," *Jurnal Belaindika: Pembelajaran dan Inovasi Pendidikan*, vol. 3, no. 2, pp. 15–19, 2021, doi: 10.52005/belaindika.v3i2.70.

[3] C. A. D. Simbolon and C. Siahaan, "Penggunaan Komunikasi Media Sosial Twitter di Kalangan Remaja di Kecamatan Cibinong, Kabupaten Bogor," *JISIP: Jurnal Ilmu Sosial dan Ilmu Politik*, vol. 10, no. 3, pp. 219–226, 2021, doi: 10.33366/jisip.v10i3.2356.

[4] Samsir *et al.*, "Naives Bayes Algorithm for Twitter Sentiment Analysis," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Jun. 2021. doi: 10.1088/1742-6596/1933/1/012019.

[5] M. K. Anam, T. P. Lestari, Latifah, M. B. Firdaus, and S. Fadli, "Analisis Kesiapan Masyarakat Pada Penerapan Smart City di Sosial Media Menggunakan SNA," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 1, pp. 69–81, 2021, doi: 10.29207/resti.v5i1.2742.

[6] K. S. Witanto, N. A. S. ER, A. E. Karyawati, I. G. A. G. A. Kadyanan, I. K. G. Suhartana, and L. gede Astuti, "Implementasi LSTM pada Analisis Sentimen Review Film Menggunakan Adam dan RMSprop Optimizer," *Jurnal Elektronik Ilmu Komputer Udayana*, vol. 10, no. 4, pp. 351–362, 2022, doi: 10.24843/JLK.2022.v10.i04.p05.

[7] V. A. Flores, L. Jasa, and L. Linawati, "Analisis Sentimen untuk Mengetahui Kelemahan dan Kelebihan Pesaing Bisnis Rumah Makan Berdasarkan Komentar Positif dan Negatif di Instagram," *Majalah Ilmiah Teknologi Elektro*, vol. 19, no. 1, p. 49, Oct. 2020, doi: 10.24843/mite.2020.v19i01.p07.

[8] Y. Zhang, J. Sun, L. Meng, and Y. Liu, "Sentiment Analysis of E-commerce Text Reviews Based on Sentiment Dictionary," in *International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, 2020, pp. 1346–1350. doi: 10.1109/ICAICA50127.2020.9182441.

[9] S. H. Lye and P. L. Teh, "Customer Intent Prediction using Sentiment Analysis Techniques," in *Proceedings of the 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 185–190. doi: 10.1109/IDAACS53288.2021.9660391.

[10] M. Abbas, K. A. Memon, A. A. Jamali, S. Memon, and A. Ahmed, "Multinomial Naive Bayes Classification Model for Sentiment Analysis," *IJCSNS International Journal of Computer Science and Network Security*, vol. 19, no. 3, p. 62, 2019, doi: 10.13140/RG.2.2.30021.40169.

[11] X. Feng, "Research of sentiment analysis based on adaboost algorithm," in *Proceedings - 2019 International Conference on Machine Learning, Big Data and Business Intelligence, MLBDBI 2019*, IEEE, 2019, pp. 279–282. doi: 10.1109/MLBDBI48998.2019.00062.

[12] 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.

[13] Junadhi, Agustin, M. Rifqi, and M. K. Anam, "Sentiment Analysis of Online Lectures using K-Nearest Neighbors based on Feature Selection," *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, vol. 11, no. 3, pp. 216–225, Dec. 2022, doi: 10.23887/janapati.v11i3.51531.

[14] M. K. Anam, B. N. Pikir, M. B. Firdaus, S. Erlinda, and Agustin, "Penerapan Naïve Bayes Classifier , K-Nearest Neighbor dan Decision Tree untuk Menganalisis Sentimen pada Interaksi Netizen dan Pemeritah," *Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, vol. 21, no. 1, pp. 139–150, 2021, doi: 10.30812/matrik.v21i1.1092.

[15] I. Mahendro and D. Abimanto, "Analisa Kepuasan Mahasiswa Terhadap E-Learning Menggunakan Algoritma Support Vector Machine," *Jurnal Saintek Maritim*, vol. 23, no. 1, pp. 97–108, 2022, doi: 10.33556/jstm.v23i1.333.

[16] F. R. Lumbanraja, R. A. Saputra, K. Muludi, A. Hijriani, and A. Junaidi, "Implementasi Support Vector Machine dalam Memprediksi Harga Rumah pada Perumahan di Kota Bandar Lampung," *Jurnal Pepadun*, vol. 2, no. 3, pp. 327–335, 2021, doi: 10.23960/pepadun.v2i3.90.

[17] N. Huda Ovirianti, M. Zarlis, and H. Mawengkang, "Support Vector Machine Using A Classification Algorithm," *Jurnal dan Penelitian Teknik Informatika*, vol. 7, no. 3, 2022, doi: 10.33395/sinkron.v7i3.

[18] I. M. Yulietha, S. A. Faraby, Adiwijaya, and W. C. Widyaningtyas, "An implementation of support vector machine on sentiment classification of movie reviews," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Apr. 2018. doi: 10.1088/1742-6596/971/1/012056.

[19] Y. N. Prasetya, D. Winarso, and Syahril, "Penerapan Lexicon Based Untuk Analisis Sentimen Pada Twiter Terhadap Isu COVID-19," *Jurnal Fasilkom*, vol. 11, no. 2, pp. 97–103, 2021, doi: 10.37859/jf.v11i2.2772.

[20] Y. Azhar, "Metode Lexicon-Learning Based Untuk Identifikasi Tweet Opini Berbahasa Indonesia," *Jurnal Nasional Pendidikan Teknik Informatika* , vol. 6, no. 3, pp. 237–243, 2017, doi: 10.23887/janapati.v6i3.11739.

[21] F. Koto and G. Y. Rahmaningtyas, "InSet Lexicon: Evaluation of a Word List forIndonesian Sentiment Analysis in Microblogs," in *International Conference on Asian Language Processing (IALP)*, Singapore: IEEE, 2017, pp. 391–394. doi: 10.1109/IALP.2017.8300625.

[22] D. Winarso, Yanda Noor Yudha, and Syahril, "Analisis Sentimen Masyarakat Pada Twiter Terhadap Isu COVID-19 Menggunakan Metode Lexicon Based," *Jurnal Fasilkom*, vol. 11, no. 2, pp. 97–103, Aug. 2021, doi: 10.37859/jf.v11i2.2772.

[23] M. K. Anam, B. N. Pikir, M. B. Firdaus, and S. Erlinda, "Penerapan Na¨ıve Bayes Classifier, K-Nearest Neighbor (KNN) dan Decision Tree untuk Menganalisis Sentimen pada Interaksi Netizen dan Pemeritah," *Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, vol. 21, no. 1, 2021, doi: 10.30812/matrik.v21i1.1092.

[24] F. F. Irfani, M. Triyanto, A. D. Hartanto, and Kusnawi, "Analisis Sentimen Review Aplikasi Ruangguru Menggunakan Algoritma Support Vector Machine," *JBMI (Jurnal Bisnis, Manajemen, dan Informatika)*, vol. 16, no. 3, pp. 258–266, 2020, doi: 10.26487/jbmi.v16i3.8607.

[25] S. Khairunnisa, A. Adiwijaya, and S. Al Faraby, "Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19)," in *Jurnal Media Informatika Budidarma*, 2021, p. 406. doi: 10.30865/mib.v5i2.2835.

[26] R. S. Putra, W. Agustin, M. K. Anam, L. Lusiana, and S. Yaakub, "The Application of Naïve Bayes Classifier Based Feature Selection on Analysis of Online Learning Sentiment in Online Media," *Jurnal Transformatika*, vol. 20, no. 1, p. 44, Jul. 2022, doi: 10.26623/transformatika.v20i1.5144.

[27] 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.

[28] A. E. Budiman and A. Widjaja, "Analisis Pengaruh Teks Preprocessing Terhadap Deteksi Plagiarisme Pada Dokumen Tugas Akhir," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 6, no. 3, Dec. 2020, doi: 10.28932/jutisi.v6i3.2892.

[29] A. A. Arifiyanti and E. D. Wahyuni, “SMOTE: Metode Penyeimbang Kelas Pada Klasifikasi Data Mining,” *SCAN*, vol. 15, no. 1, pp. 34–39, 2020, doi: 10.33005/scan.v15i1.1850.

[30] D. Ismafillah, T. Rohana, and Y. Cahyana, “Implementasi Model Support Vector Machine dan Logistic Regression Untuk Memprediksi Penyakit Stroke,” *JURIKOM (Jurnal Riset Komputer)*, vol. 10, no. 1, pp. 2407–389, 2023, doi: 10.30865/jurikom.v10i1.5478.

[31] W. P. Hutami, H. Wijayanto, and I. D. Sulvianti, “Penerapan Support Vector Machine dengan SMOTE Untuk Klasifikasi Sentimen Pemberitaan Omnibus Law Pada Situs CNNIndonesia.com,” *Xplore*, vol. 11, no. 1, pp. 26–35, 2022, doi: 10.29244/xplore.v11i1.852.