SENTIMENT ANALYSIS ON LGBT ISSUES IN INDONESIA WITH LEXICON-BASED AND SUPPORT VECTOR MACHINE ALGORITHMS

Hoiriyah[1]; Nurul Qomariya[2*)]; Aang Kisnu Darmawan[3]; Miftahul Walid[4]; Yuri Efenie[5].

[1,2,3,5] Department of Information System, [4]Department of Informatics Engineering
Universitas Islam Madura, Pamekasan, Indonesia
www.uim.ac.id
[1]hoiriyah.file.uim@gmail.com, [2*)]nurulqomariya0827@gmail.com, [3]aangdarmawan212@gmail.com,
[4]miftahwalid@gmail.com, [5]yuri.efenie.2016@gmail.com

(*) Corresponding Author

Abstract—Non-heterosexual sexual orientation (LGBT) behavior today is one of the most pervasive issues in Indonesian culture. Because of its domino effect on social stability and physical and mental health, the phenomenon known as lesbian, gay, bisexual, and transgender (LGBT) has always been under scrutiny. The development of LGBT people in Indonesia reflects cultural changes that concern many people. Freedom of speech for LGBT people on social media has many public implications. Observation of this phenomenon gives rise to views of anomalies and discrepancies that have drawn criticism. Various attempts have been made to prevent the movement of LGBT people. However, until now, many still debate the pros and cons of this LGBT movement. The lexicon-based method uses a support vector machine to classify public opinion in TikTok video comments about LGBT issues. The lexicon-based method is used as a weighting method, and the support vector machine method is used as a classification method. The results show that the highest gain in sentiment is neutral, with percentage values of 61%, 56%, 68%, 69%, and 63%. The second is positive sentiment, with percentage values of 27%, 27%, 20%, 20%, and 29%. The rest have negative sentiments. With a relatively high accuracy of the five data sets sequentially at 93%, 89%, 95%, 97%, and 91%. This shows that the majority of Indonesians prefer to ignore the issue.

Keywords: lexicon-based, LGBT, sentiment analysis, support vector machine.

Abstrak—*Perilaku orientasi seksual non-heteroseksual (LGBT) dewasa ini merupakan salah satu isu yang paling merasuk dalam budaya Indonesia. Karena efek dominonya terhadap stabilitas sosial serta kesehatan fisik dan mental, fenomena yang dikenal sebagai lesbian, gay, biseksual, dan transgender (LGBT) ini selalu mendapat sorotan. Perkembangan kaum LGBT di Indonesia mencerminkan perubahan budaya yang menjadi perhatian banyak orang. Kebebasan berbicara kaum LGBT di media sosial memiliki banyak implikasi publik. Pengamatan terhadap fenomena ini memunculkan pandangan tentang anomali dan ketidaksesuaian yang menuai kritik. Berbagai upaya telah dilakukan untuk mencegah pergerakan kaum LGBT. Namun hingga saat ini masih banyak yang memperdebatkan pro dan kontra dari gerakan LGBT ini. Metode yang digunakan berbasis leksikon dan menggunakan support vector machine untuk mengklasifikasi opini publik dalam komentar video TikTok terkait isu LGBT. Dimana metode berbasis leksikon digunakan sebagai metode pembobotan dan metode support vector machine digunakan sebagai metode klasifikasi. Hasil yang diperoleh menunjukkan bahwa perolehan sentimen tertinggi adalah sentimen netral, dengan nilai persentase 61%, 56%, 68%, 69%, dan 63%. Kedua adalah sentimen positif, dengan nilai persentase 27%, 27%, 20%, 20%, dan 29%. Sisanya memiliki sentimen negatif. Dengan akurasi yang cukup tinggi dari kelima kumpulan data secara berurutan sebesar 93%, 89%, 95%, 97%, dan 91%. Hal ini menunjukkan bahwa mayoritas masyarakat Indonesia lebih memilih untuk mengabaikan isu tersebut.*

***Kata Kunci**: analisis sentimen, lexicon-based, LGBT, support vector machine.*

INTRODUCTION

Lesbians, gays, bisexuals, and transgender people are known as LGBT. This expression has been used in place of the gay community since the 1990s because it was considered more accurate in describing the group. One of the most prominent topics today is LGBT people (Annisa & Indrawadi, 2020; Saroh & Relawati, 2017). The group that adopts the rainbow flag as its symbol has views on sexual orientation, sexual characteristics, gender identity, and gender expression that are still not widely accepted by the wider community. As a result, they often experience discrimination in the

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 19, No. 1 March 2023 | DOI: 10.33480/pilar.v18i1.4183

form of harassment, violence, threats, and other forms of harassment (Indah R. & Susilastuti, 2020; Putu Dian Adnyani, 2022). In Indonesia, LGBT people are still susceptible to issues because many groups support their right to live an everyday life based on human rights. This behavior is still debated. Quantitatively, the number of LGBT people in Indonesia is currently at a very alarming level. Several foreign and domestic independent survey institutions state that Indonesia has an LGBT population of 3% of the total Indonesian population (Azmi, 2020; Hasnah & Alang, 2019). There are several provinces with the most prominent LGBT population in Indonesia, including West Java province, with 302 thousand people detected as LBGT, and East Java province, with around 300 thousand people. Moreover, Central Java, with around 218 thousand people detected as LGBT. Moreover, DKI Jakarta, with 48 thousand people detected as ethnic LGBT (Naada, 2023; Sari, Dewi, & Morika, 2020).

Non-heterosexual sexual orientation (LGBT) behavior today is one of the most pervasive issues in Indonesian culture. Because of its domino effect on social stability and physical and mental health, the phenomenon known as lesbian, gay, bisexual, and transgender (LGBT) is always under scrutiny (Primanita, 2020). The development of LGBT people in Indonesia reflects cultural changes that concern many people. Freedom of speech for LGBT people on social media has many public implications—social media or various sources, including journals, magazines, and internet sites. Unconventional marriages, perhaps considered extreme, are often exhibited because they conflict with the terms and essence of religious teachings. The marriages mentioned above are LGBT marriages or same-sex unions. Observation of this phenomenon raises the view that it is an anomaly and a discrepancy, which has drawn criticism because it is still debatable (Gawa & Te'dang, 2023; Rahmat & Muhammad, 2023; Velando, 2020). Various attempts have been made to prevent the movement of LGBT people. However, until now, many still debate the pros and cons of this LGBT movement.

There are always LGBT cases yearly; the most recent case occurred on New Year's Eve 2023 yesterday. Bobby Nasution, Mayor of Medan, found men partnered with men (Rahmat & Muhammad, 2023). A similar case of this movement also appeared at the end of December 2022 on the tvOne news portal on YouTube yesterday with the timeline "The Movement of the LGBT Community in Garut is Starting to Worry" (Maria & Arief, 2022). Discrimination and stigma against LGBT people still exist and can affect the research results. In some cases, discriminatory language styles can be used unconsciously and undetected. Despite these problems, sentiment analysis towards the LGBT community remains essential to understand people's perceptions and how discrimination and stigma against this community develop. Therefore, a careful and concerned sentiment analysis methodology is required for the results to be accurate and useful.

Research related to LGBT has been done before. According to (Fitri, Andreswari, & Hasibuan, 2019), in analyzing sentiments related to LGBT, the naïve Bayes method is more suitable than the random forest or decision tree method. This is because the accuracy value of the naïve Bayes method is 83.43%. More powerful than the random forest and decision tree methods, which only have an accuracy value of 82.91%. This shows that the Naïve Bayes method is more suitable for sentiment analysis than the other two. This statement is reinforced by research conducted by (Anjani, Chamid, & Murti, 2022), which shows that research related to sentiment analysis using the Naïve Bayes method obtains an accuracy of 95% with a positive sentiment of 77% and a negative 150.2%. Another study conducted by (Aldinata, Soesanto, Chandra, & Suhartono, 2023) showed that of the five methods used in comparing opinions from Twitter, the logistic regression method had the highest score with a precision of 0.7233, recall of 0.7006, and F-1 of 0.7087. This proves that logistic regression is more suitable among the five methods used in the study. Recent research was also conducted by (Ardras & Voutama, 2023) regarding anti-LGBT campaigns on Twitter social media using the Naïve Bayes method on rapid miners with a data ratio of 60:40, and the results show that 55% of the data classified as neutral sentiment and 45% are in support of anti-LGBT, while negative perceptions arise due to public disappointment with anti-LGBT.

Based on previous research, there has been no research on LGBT issues based on data from TikTok. As a platform, TikTok is more open to people posting queer content, such as applying makeup to a transgender person or witnessing two men embracing it. The videos in the space are uplifting and even brutally honest, despite concerns about suppressing LGBTQ content (Romadlon, Ayuningtyas, & Sundayani, 2022). In previous studies, the average data was from Twitter and social media. So in this study, research will be carried out on "sentiment analysis of LGBT issues in Indonesia using lexicon-based methods and support vector machines," where the data used comes from video comments related to LGBT issues on TikTok taken at five different times. The data will be processed using the lexicon-based method as the weighting method and the support vector machine as the classification method. The purpose of this

research is to apply an algorithm based on Lexicon and a Support Vector Machine (LB-SVM) to analyze community comments in TikTok videos about LGBT issues in Indonesia and to test the performance of an algorithm based on Lexicon and a Support Vector Machine (LB-SVM). It is hoped that the results of this study can reveal people's opinions about whether or not they support LGBT issues in Indonesia so that the results are expected to be used as material for consideration by policymakers in implementing better and more targeted policies to address issues related to LGBT people in Indonesia.

MATERIALS AND METHODS

The lexicon-based technique and the support vector machine were employed in this study. This is because it is based on a literature review (Cindo & Rini, 2019) and (Singgalen, 2021). This study revealed that sentiment analysis research most frequently employed the support vector machine and naive Bayes approach. However, logistic regression and lexicon-based approaches get the best results. As a result, this study employed the support vector machine approach as a classification method, and the lexicon-based method was used as a weighting method. The design of the algorithm proposed in this study is as follows:

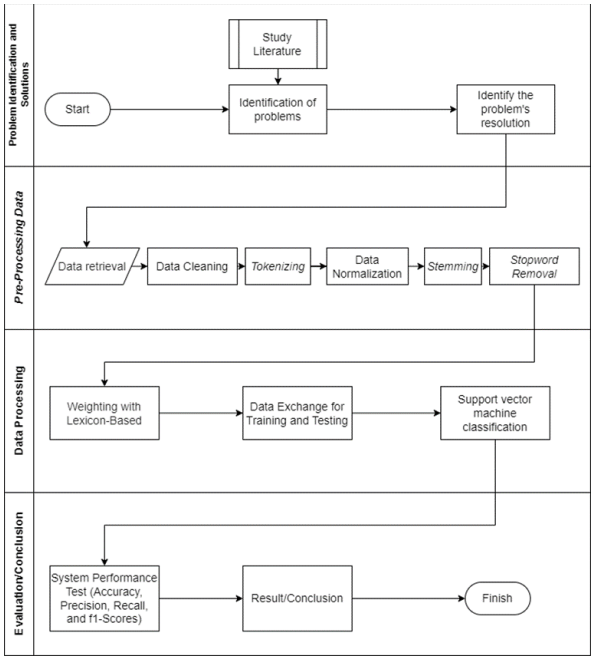

Figure 1. Research Stages

In Figure 1, you can see the stages passed in this study.

1. Problem Identification and Solutions

Identification of the problem is done by looking at previous studies that have been done. The search used the keyword "sentiment analysis LGBT on several journal websites such as ieeexplore.ieee.org, sciencedirect.com, and garuda.kemdikbud.co.id. After finding the topic of the problem to be raised, a search for the method to solve the problem is carried out.

2. Pre-processing Data

Pre-processing data is the initial stage in the text analysis process, which aims to clean and change the raw data format into a more structured format ready for analysis. In pre-processing data, several steps must be followed, including:

a. Data Retrieval

The data used in this study was taken from video comments on TikTok related to LGBT issues. Data was obtained by scraping using an application (Kubernetes, 2021/2023) with the help of inspecting elements in the browser with the keyword "LGBT Indonesia". Data was taken in stages from February 22, 2023, to May 18, 2023, and then analyzed separately.

b. Data Cleaning

Data cleaning is done to eliminate differences that exist in opinion data. Things to do in this stage include removing numbers, blank spaces, hyphens, punctuation marks, excess letters, hyperlinks, and case folding. Case folding is a stage in which words are changed into the same form. For example, from uppercase to all lowercase, or vice versa. After case folding, data cleaning removes punctuation marks, numbers, links, mentions, hashtags, etc. After the data cleaning, data replacement will be done, namely removing excess letters in sentences.

c. Tokenizing

Tokenizing is the process of dividing sentences into parts called tokens. Tokens can be formed in words, phrases, or other meaningful elements. In this stage, the text or opinion sentences that have been cleaned will be broken down into word chunks; these word chunks are called tokens or terms. At this stage, the tokenizing process is carried out with the help of NLTK (natural language toolkit), a Python library focusing on text data processing.

d. Data Normalization

At this stage, changes to slang sentences are made into standard forms. After the data is cleaned, lang slang words are changed to a standard form to make the analyzed text more accurate. Next, normalization is done, namely changing the word to a standard form.

e. Stemming

Stemming is done by changing words into essential words according to the word structure. This stage is done by removing the affixes in the word to get the original word. This stage is carried out with the help of a Python library, namely StemmerFactory from Sastrawi and Swifter, to

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 19, No. 1 March 2023 | DOI: 10.33480/pilar.v18i1.4183

speed up the steaming process. In the process, the normalized data is changed into essential words. Words that have been divided into tokens are checked in the primary dictionary. If the token is already a base word, it will be stored for the following process. However, if the token is not included in the root word, cleaning will be carried out by removing the affixed word so that it forms a base word.

f. Stopword Removal

This stage eliminates words that are less meaningful and do not contain any sentiment but appear frequently in the text. After the data is made into essential words in the stemming stage, stopword checking (common words) is carried out. The token will be deleted if included in the general word. Then update the sentences in the data. However, if the token is not included in the general world, it will be stored in the data list according to the provisions.

3. Data Processing

After pre-processing the data, the next step is data processing. Data processing is carried out using two methods, including:

a. Weighting with Lexicon-Based

Weighting is done based on a dictionary or lexicon. The calculation process begins by looking for lexicon values based on the dictionary. After knowing which words are positive, negative and neutral in a sentence, calculate each word containing sentiment by adding up the value opinion. The sum of the opinion values for good sentiment is worth one or more, while neutral in sentences for those with a price = 0, and vice versa for negative in sentences with a very high value = -1 (Mahendrajaya, Buntoro, & Setyawan, 2019; Oktaviana, Sari, & Indriati, 2022).

The first step in data processing is weighting. Weighting is done based on the dictionary (lexicon). Before weighting, the data was translated from Indonesian into English to simplify the weighting process. Translating data is done with the help of the Python library, namely Translate from GoogleTrans. After the data is successfully translated, it is weighted using the lexicon-based method by utilizing the SentimentIntensityAnalyzer library from VaderSentiment in Python.

b. Data Exchange for Training and Testing

The next step is data classification after the data has been successfully weighted. However, before classifying the data, it is divided into training and test data. This data will later be used to train and test a data classification model. Data sharing is done by utilizing the train_test_split library from sci-kit-learn.model_selection in Python programming with training data sharing of 80% and test data sharing of 20%.

c. Support Vector Machine Classification

The method finds the best hyperplane by maximizing the distance between classes. Hyperplane is a function that is used to separate two classes and is used to classify higher dimensional classes. The support vector machine uses a kernel trick to convert the data to higher dimensions to separate the data linearly (Oktaviana et al., 2022).

The data classification process was carried out in this study using the support vector machine method. The classification model is done with the help of the Python library, namely SVM.SVC from sci-kit-learn. Data given weight is entered into the model and trained with training data that has been divided before. After that, the model will validate the data with test data to get predictions from the classification results.

4. Evaluation/Conclusion

After the data is classified, an evaluation of accuracy, recall, and precision is then carried out to determine the accuracy of the algorithm used in classifying the data. At this stage, an evaluation of the performance of the model is carried out. Tests are carried out to determine the system's success level in testing the data used. Testing was carried out using the Python library, namely classification_report from sci-kit-learn. classification_report is a performance evaluation report of a classification model. This report summarizes several metrics, such as accuracy, precision, recall, and f1_score. The classification_report compares the predictions generated by the model with the actual data.

RESULTS AND DISCUSSION

1. Results of problem identification and solutions

In identifying the previously identified problems, seven studies were found related to sentiment analysis on LGBT issues. Five of the seven studies used data from Twitter, while two used news sites. This shows that from previous research, there has been no research related to LGBT issues based on data from TikTok.

From previous research, there has also been no related research that examines LGBT people using a combination of lexicon-based methods and support vector machines. So in this study, the method is used to solve existing problems.

2. Results of pre-processing data

a. Results of data retrieval

The data is from commentary from several TikTok videos with LGBT topics. Data collection was carried out in five stages: on February 22, 2023, 877 data were obtained. On February 28, 2023, 3,208 data were obtained. On May 2, 2023, 5,989 data were obtained. On May 11, 2023, 4,748 data were

obtained; on May 18, 2023, as many as 3267 comments were obtained.

b. Results of data cleaning

After the data has been successfully retrieved, the next step is data cleaning, which removes differences in the data. The following is the result of the data-cleaning process:

Table 1. Results of data cleaning

CommentText	*CaseFolding*	*Cleaning*
gk gk gw pokoknya harus bisa berhentiin ini	gk gk gw pokoknya harus bisa berhentiin ini	gk gk gw pokoknya harus bisa berhentiin ini
NKRI MENGHARGAI PERBEDAAN BUKAN PENYIMPANGAN ðŸ™ðŸ™ðŸ™	nkri menghargai perbedaan bukan penyimpangan ðŸ™ðŸ™ðŸ™	nkri menghargai perbedaan bukan penyimpangan
astagfirullah astagfirullah matakuuuuu yaampun ya Allah gusti	astagfirullah astagfirullah matakuuuuu yaampun ya allah gusti	astagfirullah astagfirullah matakuu yaampun ya allah gusti

In Table 1, it can be seen the results of the cleaning process where several words that have capital letters have been successfully folded into all lowercase letters. After the case folding process, you can also see the replace process results, where words with more than two of the same letters are deleted, leaving two of the same letters. After that, a cleaning process is carried out where inappropriate characters such as @, #, %, and so on are removed to facilitate the data pre-processing process. After the data cleaning process, irrelevant variables from the data were removed to improve the quality and accuracy of the analysis. Writing errors and inaccurate formatting are corrected so that it becomes a dataset that is easier to interpret.

c. Results of Tokenizing

After the data is cleaned, tokenization is carried out, where the cleaned data is separated into tokens. The result is as follows in Table 2:

Table 2. Results of tokenizing

Cleaning	*Tokenize*
gk gk gw pokoknya harus bisa berhentiin ini	['gk', 'gk', 'gw', 'pokoknya', 'harus', 'bisa', 'berhentiin', 'ini']
nkri menghargai perbedaan bukan penyimpangan	['nkri', 'menghargai', 'perbedaan', 'bukan', 'penyimpangan']
astagfirullah astagfirullah matakuu yaampun ya allah gusti	['astagfirullah', 'astagfirullah', 'matakuu', 'yaampun', 'ya', 'allah', 'gusti']

In Table 2, it can be seen the results of the tokenizing process where the sentences that have been processed at the cleaning stage are successfully separated by spaces. Each row in the data is converted to a token. These tokens will later be processed individually in the data normalization process. This process is essential in data processing so that analysis can be carried out in more detail.

d. Results of data normalization

The next stage is data normalization. At this stage, the data in slang or abbreviations is changed back to the word's original form. The following are the results of data normalization:

Table 3. Results of data normalization

Tokenize	Normalisasi data
['gk', 'gk', 'gw', 'pokoknya', 'harus', 'bisa', 'berhentiin', 'ini']	['tidak', 'tidak', 'saya', 'pokoknya', 'harus', 'bisa', 'menghentikan', 'ini']
['nkri', 'menghargai', 'perbedaan', 'bukan', 'penyimpangan']	['nkri', 'menghargai', 'perbedaan', 'bukan', 'penyimpangan']
['astagfirullah', 'astagfirullah', 'matakuu', 'yaampun', 'ya', 'allah', 'gusti']	['astaghfirullah', 'astaghfirullah', 'mataku', 'ya ampun', 'iya', 'allah', 'gusti']

Table 3 shows the results of data normalization, where slang words and abbreviations are changed to their original form. For example, as shown in the table of the first six rows, the words "gk", "me", and "stop" were successfully changed to the words "no", "me", and "stop". This is done to facilitate the steaming process.

e. Results of stemming

The next stage is stemming, where the normalized words are converted into essential words. The results of this stage can be seen in the following table 4:

Table 4. Results of stemming

Normalisasi data	*Stemming*
['tidak', 'tidak', 'saya', 'pokoknya', 'harus', 'bisa', 'menghentikan', 'ini']	['tidak', 'tidak', 'saya', 'pokok', 'harus', 'bisa', 'henti', 'ini']
['nkri', 'menghargai', 'perbedaan', 'bukan', 'penyimpangan']	['nkri', 'harga', 'beda', 'bukan', 'simpang']
['astaghfirullah', 'astaghfirullah',	['astaghfirullah', 'astaghfirullah',

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 19, No. 1 March 2023 | DOI: 10.33480/pilar.v18i1.4183

Normalisasi data	*Stemming*
'mataku', 'ya ampun', 'iya', 'allah', 'gusti']	'mata', 'ya ampun', 'iya', 'allah', 'gusti']

In Table 4, you can see the results of the stemming, where the results of the normalized words have been successfully converted into elemental forms. The stemming process works by removing prefixes and suffixes from words. In Table 4, it can be seen that several words such as "in essence", "stop", "appreciate", "difference," and others have changed into primary word forms, namely "principal", "stop", "price," and "difference". This is done to simplify the weighting process.

f. Results of stopword removal

The final stage of data pre-processing is stopword removal. At this stage, words that do not provide helpful information in a sentence are removed. The results can be seen in the following table 5:

Table 5. Results of stopword removal

Stemming	*Stopword Removal*
['tidak', 'tidak', 'saya', 'pokok', 'harus', 'bisa', 'henti', 'ini']	['pokok henti']
['nkri', 'harga', 'beda', 'bukan', 'simpang']	['nkri harga beda simpang']
['astaghfirullah', 'astaghfirullah', 'mata', 'ya ampun', 'iya', 'allah', 'gusti']	['astaghfirullah astaghfirullah mata ampun allah gusti']

In Table 5, you can see the results of the stopword removal process, where unnecessary words in a sentence were successfully deleted. It can be seen in the table of the first five rows that words such as "no", "I", "should", "could," and "this" were successfully removed at this stage. This is done to facilitate the weighting process.

3. Results of data processing

a. Results of weighting with lexicon-based

The first stage in data processing is weighting. Before weighting is done, it is first translated into English to facilitate weighting. After that, the data is weighted using the lexicon-based method.

Table 6. Results of weighting with lexicon-based

Stopword Removal	Compound Score
['stop tree']	-0.2960
['nkri price different intersection']	0.0000
['astaghfirullah astaghfirullah eye mercy allah gusti']	0.3612

Table 6 shows the results of the weighting with the lexicon-based method. In the weighting, there are three weights: a value less than 0 is defined as a negative sentiment, a value of 0 is a neutral sentiment, and a value above 0 is defined as a positive sentiment.

b. Results of data exchange for training and testing

The second stage in data processing is the distribution of training data and test data in a ratio of 80:20. The divided data contains four tables, with the first three tables serving as training data and the remaining 1 table as test data. Following are the results of the distribution of training data and test data:

Table 7. Results of data exchange for training and testing

Data	Data Latih	Data Uji	Jumlah
22-02-23	428	107	535
28-02-23	1557	390	1947
02-05-23	1667	417	2084
11-05-23	1657	415	2072
18-05-23	1390	384	1738

In Table 7, the distribution of training data and test data for the five data sets used can be seen. In the first data set, of the 535 data points used, 428 were used as training data, and 107 were used as test data. In the second set of data, of the 1947 data used, 1557 were used as training data, and 390 were used as test data. In the third data set, of the 2084 data points used, 1667 were used as training data, and the remaining 417 were used as test data. In the fourth data set, of the 2072 data points used, 1657 were used as training data, and 415 were used as test data. Whereas in the fifth data set, of the 1738 data points used, 1390 were used as training data and 384 as test data.

c. Results of support vector machine classification

The next stage is to conduct model training on the training data that has been previously divided. After that, the model validation process is carried out using test data that has been divided previously to produce a prediction of data classification. Following are the results of data classification using the support vector machine method:

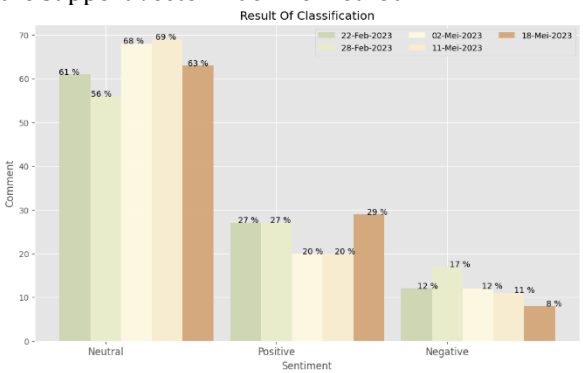

Figure 2. Results of support vector machine classification

Figure 2 shows that the results of classifying public opinion on social media Tiktok with LGBT issues tend to have a neutral opinion. The graph above shows that in the five data sets tested, neutral sentiment has the most value, namely 61%, 56%, 68%, 69%, and 63%, respectively. At the same time, the least sentiment is negative sentiment, with respective values of 12%, 17%, 12%, 11%, and 8% in the five data points. This shows that public opinion on social media Tiktok tends to have neutral opinions. Only a few have opposing opinions on LGBT issues.

4. Results of Evaluation/Conclusion

After the data has been classified, the model performance test is carried out to determine the success rate of the model that has been made.

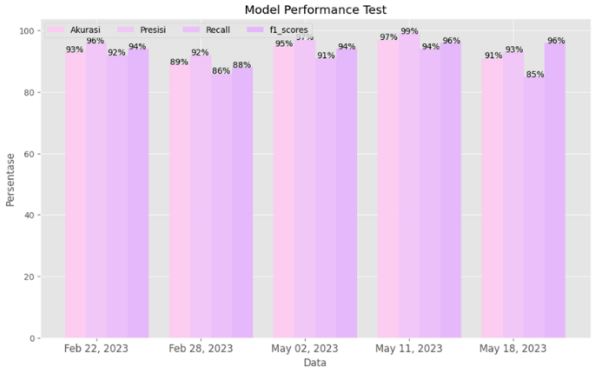

Figure 3. Results Of Model Performance Test

The results of the model testing of the five data sets are shown in Figure 3. As a result, the data taken on February 22, 2023, after being tested, obtained an accuracy value of 93%, 92% recall, 96% precision, and 94% f1_scores. On February 28, 2023, an accuracy of 89%, 86% recall, 92% precision, and 88% f1_scores was obtained. Data taken on May 2, 2023, obtained an accuracy of 95%, a recall of 91%, a precision of 97%, and f1_scores of 94%. Data taken on May 11, 2023, obtained results of 97% accuracy, 99% precision, 94% recall, and 96% f1-scores, and data taken on May 18, 2023, obtained results of 91% accuracy, 93% precision, 85% recall, and 96% f1-scores.

5. Discussion

This research was conducted because, until now, LGBT is still a sensitive issue among the public. There are still many pros and cons regarding the LGBT movement that has taken place in Indonesia. The results of this study are expected to be able to express a public opinion whether they support, reject or are neutral towards the LGBT issue in Indonesia. Thus, it is anticipated that the findings of this study will be utilized as material for policymakers to consider when establishing better and more specific policies to address concerns linked to LGBT people. The data used in this study comes from the social media TikTok. This is because, in previous studies related to LGBT, no one has used data derived from video comments from Tiktok. Data were obtained at five different times and analyzed separately according to the time of data collection. This is done to see how the development of public opinion in dealing with the LGBT issue is happening in Indonesia. The data is then processed using the lexicon-based weighting method and support vector machine as a classification method.

The results of this study indicate that public opinion on LGBT issues tends to be neutral. This can be seen in Figure 2, where the graph shows that of the three data taken, the highest results are in the neutral category, with the acquisition of neutral sentiment of on February 22, 2023, 61%, and on February 28, 2023, of 56%, on May 2, 2023, obtained a sentiment yield of 68%, on May 11, 2023, obtained a yield of 69%, and on May 18, 2023, obtained a sentiment yield of 63%. As for negative sentiment of 12%, 17%, 12%, 11% and 8% for the five data. And for the positive sentiment of 27%, 27%, 20%, 20% and 29% for the five data, respectively. These results show that people are increasingly not showing positive and negative attitudes towards LGBT issues. They choose to be neutral. After a classification process is carried out, a model performance test is carried out so that the results obtained can be seen in Figure 3, where the graph shows the accuracy of the model used in classifying data using the support vector machine method, which is relatively high and accurate, namely 93%, 89%, 95%, 97% and 91% of the five data respectively. From these results, it can also be seen that the data taken on May 11, 2023, obtained the best accuracy results. However, literature and language experts have not checked this study manually, so the truth cannot be ascertained.

The results obtained in this study were influenced by how the data was weighted. This is because the weighting is done automatically using a library available in the Python programming language. Someone's opinion sometimes differs from another's, especially opinions processed automatically by machines. In future research, it is hoped that some experts in literature and language can carry out manual weighting to find out the actual context of language.

In previous studies, research was conducted regarding the anti-LGBT campaign by (Ardras & Voutama, 2023) on Twitter social media using the Naive Bayes method on rapid miners with a data ratio of 60:40. The results showed that 55% of the data classified had neutral sentiment and 45% supported anti-LGBT, while negative perceptions arise due to public disappointment with anti-LGBT.

While the research currently being carried out collects data with the keyword Indonesian LGBT on social media Tiktok, which is processed using the lexicon-based method and a support vector machine using the Python programming language with a comparison of training data and test data of 80:20 for each data set tested.

CONCLUSION

Sentiment analysis using the lexicon-based method and support vector machine on LGBT issues from TikTok video comments obtained an accuracy of 93%, 89%, 95%, 97%, and 91%, respectively. This is based on testing the accuracy of the TikTok video comment data, which was previously taken regularly at five different times. Most sentiment results are neutral, with percentage values of 61%, 56%, 68%, 69%, and 63%. The second is positive sentiment, with percentage values of 27%, 27%, 20%, 20%, and 29%. The rest have negative sentiments. This shows that the majority of Indonesians prefer to ignore the issue. However, not a few also have a positive attitude towards the LGBT issues that occur. So it is hoped that the results of this research can assist policymakers in addressing the opinions of the Indonesian people regarding current LGBT issues. This research still has drawbacks because, in the process, it was only carried out with the help of a system without manual processing by linguists and literature experts. Another drawback is that the data collection process is still done manually and is only taken at five different times. So that in future research, data collection can be carried out automatically with a more extended period and a more significant amount of data and tested with different classification methods such as naive Bayes, random forest, and others.

REFERENCE

Aldinata, Soesanto, A. M., Chandra, V. C., & Suhartono, D. (2023). Sentiments comparison on Twitter about LGBT. *Procedia Computer Science*, *216*, 765–773. https://doi.org/10.1016/j.procs.2022.12.194

Anjani, A. M., Chamid, A. A., & Murti, A. C. (2022). Analisis Sentimen Kaum Lgbt Pada Media Sosial Twitter Menggunakan Algoritma Naïve Bayes. *Jurnal Teknik Informatika*, *1*(2), 1–8. https://doi.org/10.02220/jtinfo.v1i2.259

Annisa, O., & Indrawadi, J. (2020). Peran Pemerintah dalam Menanggulangi LGBT di Kota Payakumbuh. *Journal of Civic Education*, *3*(1), 110–118. https://doi.org/10.24036/jce.v3i1.341

Ardras, D. W., & Voutama, A. (2023). Analisis Sentimen Anti Lgbt Di Indonesia Melalui Media Sosial Twitter. *Jurnal Teknika*, *15*(1), 23–28. https://doi.org/10.30736/jt.v15i1.926

Azmi, K. R. (2020). Student's LGBT Trend Analysis of Transgender Counseling Through WOCA (Wisdom-Oriented Counseling Approach). *Prophetic: Professional, Empathy and Islamic Counseling Journal*, *3*(1), 25–36. http://dx.doi.org/10.24235/prophetic.v3i1

Cindo, M., & Rini, D. P. (2019). Literatur Review: Metode Klasifikasi Pada Sentimen Analisis. *Seminar Nasional Teknologi Komputer & Sains (SAINTEKS)*, 66–70. Retrieved from https://seminar-id.com/prosiding/index.php/sainteks/article/view/124

Cubernetes. (2023). *Functionality* [Python]. Retrieved from https://github.com/cubernetes/TikTokCommentScraper (Original work published 2021)

Fitri, V. A., Andreswari, R., & Hasibuan, M. A. (2019). Sentiment Analysis of Social Media Twitter with Case of Anti-LGBT Campaign in Indonesia using Naïve Bayes, Decision Tree, and Random Forest Algorithm. *Procedia Computer Science*, *161*, 765–772. https://doi.org/10.1016/j.procs.2019.11.181

Gawa, E. C. S., & Te'dang, V. (2023). Penggunaan Media Sosial Sebagai Simbol dalam Mendukung Hubungan LGBT. *Journal on Education*, *05*(04), 15598–15608. https://doi.org/10.31004/joe.v5i4.2669

Hasnah, H., & Alang, S. (2019). Lesbian, Gay, Biseksual dan Transgender (Lgbt) Versus Kesehata: Studi Etnografi. *Jurnal Kesehatan*, *12*(1), 63–72. https://doi.org/10.24252/kesehatan.v12i1.9219

Indah R., A., & Susilastuti, D. H. (2020). Different Types Of Stereotype Toward Lgbt As Minority On American Online News. *Rubikon : Journal of Transnational American Studies*, *7*(1), 35–46. https://doi.org/10.22146/rubikon.v7i1.62510

Mahendrajaya, R., Buntoro, G. A., & Setyawan, M. B. (2019). Analisis Sentimen Pengguna Gopay Menggunakan Metode Lexicon Based dan Support Vector Machine. *Komputek*, *3*(2), 52–63. https://doi.org/10.24269/jkt.v3i2.270

Maria, A., & Arief, F. (Directors). (2022). *Gerakan Komunitas LGBT di Garut Mulai Mengkhawatirkan | Apa Kabar Indonesia Pagi tvOne*. Retrieved from https://www.youtube.com/watch?v=0yvU__KEi6g

Naada, N. (2023). Lima Daerah Di Indonesia Dengan Populasi LGBT Terbanyak. Retrieved March 25, 2023, from Medialokal.co website: https://medialokal.co/news/detail/42644/lima-daerah-di-indonesia-dengan-populasi-lgbt-terbanyak

Oktaviana, N. E., Sari, Y. A., & Indriati, I. (2022). Analisis Sentimen terhadap Kebijakan Kuliah Daring Selama Pandemi Menggunakan Pendekatan Lexicon Based Features dan Support Vector Machine. *Jurnal Teknologi Informasi dan Ilmu Komputer*, *9*(2), 357–362. https://doi.org/10.25126/jtiik.2022925625

Primanita, Y. (2020). Emotional Quotient dan Perilaku Self Injury pada LGBT. *Jurnal RAP (Riset Aktual Psikologi Universitas Negeri Padang)*, *11*(1), 90–103. https://doi.org/10.24036/rapun.v11i1.109779

Putu Dian Adnyani. (2022). Problematika Perlindungan Hukum Terhadap Kelompok Lesbian, Gay, Biseksual, dan Transgender (Lgbt) dalam Perspektif Ham Internasional. *Ganesha Law Review*, *4*(1), 35–44. https://doi.org/10.23887/glr.v4i1.1501

Rahmat, U., & Muhammad, R. (2023). Penegasan Bobby Usai Lihat Pasangan Cowok-Cowok: Kota Medan Anti-LGBT. Retrieved March 25, 2023, from Kumparan website: https://kumparan.com/kumparannews/penegasan-bobby-usai-lihat-pasangan-cowok-cowok-kota-medan-anti-lgbt-1zYhTJRUtb9

Romadlon, F. N., Ayuningtyas, G., & Sundayani, R. (2022). The Effect Of Lgbt's Content In Tik Tok On The Acceptance Of Lgbt By College Students In A Campus Environment. *Makna: Jurnal Kajian Komunikasi, Bahasa, Dan Budaya*, *10*(1), 50–58. https://doi.org/10.33558/makna.v10i1.2494

Sari, I. K., Dewi, R. I. S., & Morika, H. D. (2020). Bahaya Lesbian, Gay, Biseksual, Transgenders (Lgbt) Di Sma Kota Padang. *Jurnal Abdimas Saintika*, *2*(1), 85–90. http://dx.doi.org/10.30633/jas.v2i1.570

Saroh, Y., & Relawati, M. (2017). Indonesian Youth's Perespective Towards Lgbt. *Humanus*, *16*(1), 71. https://doi.org/10.24036/jh.v16i1.7323

Singgalen, Y. A. (2021). Pemilihan Metode dan Algoritma dalam Analisis Sentimen di Media Sosial: Sistematic Literature Review. *Journal of Information Systems and Informatics*, *3*(2), 278–302. https://doi.org/10.33557/journalisi.v3i2.125

Velando, M. (2020). Analysis of Constitutional Court Decision Number 46/PUU-XIV/2016 Related to LGBT and Community Attitude. *Journal of Law and Legal Reform*, *1*(2), 259–272. https://doi.org/10.15294/jllr.v1i2.35767

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 19, No. 1 March 2023 | DOI: 10.33480/pilar.v18i1.4183

This sheet is intentionally left blank