Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 3, July 2024 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i3.13733 p-ISSN : 2541-044X

Combination of Lexical Resources and Support Vector Machine for Film Sentiment Analysis

Putri Agustina [1)], Raissa Amanda Putri [2)]
[1)]Sains dan Teknologi,Ilmu Komputer, Universitas Islam Negeri Sumatera Utara, Medan, Indonesia
[2)]Sains dan Teknologi, Universitas Islam Negeri Sumatera Utara, Jl. Lap. Golf No. 120, Medan, Indonesia
[1)]putri310802@gmail.com, [2)]raissa.ap@uinsu.ac.id

Submitted : Jun 7, 2024 | **Accepted** : Jun 25, 2024 | **Published** : Jul 1, 2024

Abstract: Text data generated by internet users holds potentially valuable information that can be researched for new insights. One strategy for obtaining information from a text data set is to classify text into predetermined categories based on existing data. Text classification is an aspect of Text Mining. One of the popular approaches in Text Mining uses the Support Vector Machine (SVM) classification algorithm, which aims to classify text and separate data into different classes. However, in some cases, SVM classification algorithms may face difficulties in understanding the context of the text properly due to unclear wording, varying sentence structures, or a lack of understanding of interpretation. To address this problem, applying SVM classification using lexical resources can be an effective solution. In this research framework, the first step is to obtain data, which in this case is a film review dataset taken from the kaggle.com site. After obtaining the data, the next step is preprocessing. The results of the preprocessing are then divided into 80:20 percentages. The 80% training data is used to search for the form of polarization, and this training data lexicon is used for training the SVM model. Based on the modeling results, the overall model accuracy is around 85%, calculated using the confusion matrix. The precision value, which shows the proportion of correct positive predictions, reached 88%. The precision for negative predictions reached 80%, and for neutral predictions, it reached 0%. These results show that the Lexicon+SVM model has good performance, with an accuracy of 85%.

Keywords: Text Mining; Support Vector Machine; Lexical Resources; Sentiment; Combination

INTRODUCTION

In the increasingly advanced digital era of information, the amount of text data produced daily is rapidly increasing. Text data generated by internet users includes reviews, film comments, social media comments, and so on. This data holds potentially valuable information that can be researched for new insights. However, manual processing and analysis of large and complex text data require significant time. Therefore, an effective method is needed to process and analyze text data.

Text Mining is a method that is rapidly developing in text data analysis. Text Mining is the process of extracting information or identifying current issues using mechanisms to analyze large amounts of data (Jo, 2019). Text Mining combines different parts of text based on certain rules to extract information from unstructured text (Kusnia et al., 2022). One strategy for obtaining information from a set of text

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 3, July 2024 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i3.13733 p-ISSN : 2541-044X

data is to classify the text into predetermined categories based on existing data. Text classification is an aspect of Text Mining that involves predicting class categories from data (Rahman et al., 2021).

Classification algorithms capable of performing text mining include the Support Vector Machine (SVM). Support Vector Machine is one of the machine learning algorithms used for classification and regression. Support Vector Machine (SVM) was further developed by Boser, Guyon, and Vapnik, and was first presented in 1992 at the Annual Workshop on Computational Learning Theory. This algorithm is also considered one of the best due to its pattern recognition capabilities, achieved by transforming data in the input space to a higher dimensional space (feature space) and performing optimization in the new vector space (Utama et al., 2019).

One of the popular approaches in Text Mining uses the Support Vector Machine (SVM) classification algorithm to classify text and separate data into different classes. However, in some cases, SVM classification algorithms may face difficulties in properly understanding the context of the text due to unclear wording, varying sentence structures, or lack of interpretative understanding. To address this problem, applying SVM classification using lexical resources can be an effective solution. This involves using lexical resources, such as dictionaries of words with specific polarities, to classify texts into positive, negative, or neutral categories based on the polarity of the words contained in the text. The process involves parsing text into individual words and then pairing these words with values (Hofmann & Chisholm, 2016). By using lexical resource-based Text Mining techniques, the results of Text Mining and SVM learning classify text into positive, negative, or neutral categories based on the polarity of the words found. This method allows for easier interpretation and provides a better understanding of the views or opinions contained in texts, such as film reviews, film comments, or news articles.

Based on the description above, this research develops a Text Mining method based on the use of lexical resources for training the SVM classification algorithm. It involves studying the selection and integration of appropriate lexical resources, utilizing these resources to address unclear words in text, and evaluating the performance of the Text Mining method based on lexical resources. The results of this research can be useful in various applications such as document classification and topic identification in text.

LITERATURE REVIEW

Text Mining

Text Mining is a part of data mining that has the same goal: mining data to find unique patterns or relationships that represent the content or special characteristics of a text document (Nanda et al., 2022). Text Mining is a method for extracting information from text to identify patterns or trends contained therein (Firdaus & Firdaus, 2021). It focuses on text analysis, which is considered to have higher commercial value than data mining itself (Firdaus & Firdaus, 2021). Text Mining is a comprehensive and relevant research area in almost every aspect of our lives (Alhaq et al., 2021). It seeks to find patterns or relationships between data in documents (Firdaus & Firdaus, 2021). This technique aims to reveal new information or trends by processing and analyzing large amounts of data that may have previously gone undetected. Text obtained from social media, which generally does not use standard language, requires processing using Text Mining. The Text Mining process involves various stages, such as text categorization, text clustering, concept/entity extraction, granular taxonomy creation, sentiment analysis, document inference, and entity relationship modeling (Utama et al., 2019). The process of preparing document text or raw datasets is also called text preprocessing. Text preprocessing functions to convert unstructured or random text data into structured data (Furqan et al., 2022).

Lexical Resources

A lexical resource or lexicon-based dataset is a word dictionary used to carry out sentiment analysis (Hoiriyah et al., 2023). Lexical resources are collections that contain linguistic information about words and other linguistic entities. They can be of various types, such as dictionaries, thesauruses, text corpora, word embeddings, gazetteers, stopword lists, and so on. These dictionaries contain words in a particular language and assign values or weights to specified phrases. The weighting of each word can be negative, neutral, or positive (Hoiriyah et al., 2023). The intensity of emotion greatly influences the strength of sentiment. For example, the emotion "very happy" reflects a very positive sentiment, while "very angry" reflects a very negative sentiment. Analyzing emotions in texts, such as news or social media posts, is

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 3, July 2024 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i3.13733 p-ISSN : 2541-044X

useful for understanding users' emotional responses to certain conversation topics and helps in understanding how emotions influence human behavior (Aribowo & Khomsah, 2021). The aim of using lexical resources in natural language processing and text mining is to improve the understanding and analysis of text by utilizing available linguistic information. This facilitates better and deeper processing, understanding, and analysis of text in various natural language processing and text mining applications (Scutelnicu, 2023).

There are three methods to collect sentiment lexicons: the manual approach, the dictionary-based approach, and the corpus-based approach (Hamka & Ratna Sari, 2022). Although lexicon-based approaches generally have a lower level of accuracy compared to corpus-based approaches, their reliability is highly dependent on the quality and amount of training data available (Saputra et al., 2021). Examples of lexical resources include the Indonesian Sentiment Lexicon, VADER, AFINN, Hu Liu Lexicon, and SentiWordNet. Each lexical resource has different characteristics. For example, VADER considers punctuation and capitalization to increase the accuracy of sentiment values in a text (Hayaty & Pratama, 2023).

Support Vector Machine

One approach that has recently received significant attention in pattern recognition is the Support Vector Machine (SVM). SVM, developed by Boser, Guyon, and Vapnik, was first introduced in 1992 at the Annual Workshop on Computational Learning Theory (Putri et al., 2020). Unlike neural network strategies that try to find a separating hyperplane between classes, SVM looks for the best hyperplane in the input space. The basic principle of SVM is a linear classifier, which is then extended to handle non-linear problems (Putri et al., 2020). In SVM, situations often involve data that is not linearly separable (Kavabilla et al., 2023). To overcome this, Kernel techniques are used to solve non-linear problems. This approach involves mapping data from a low-dimensional space to a higher-dimensional space, often referred to as feature space (Kavabilla et al., 2023). A hyperplane is a function that separates two classes in a high-dimensional space. SVM uses Kernel tricks to transform data into a higher-dimensional space so that it can be separated linearly (Oktaviana et al., 2022) notation below.

$$K(x_i, x_j) = \phi(x_j).\phi(x_j) \qquad (1)$$

Some commonly used Kernel functions include Linear, Polynomial, and Radial Basis Function (RBF). One type of Kernel that is often used in Support Vector Machines (SVM) is the Radial Basis Function (RBF) Kernel (Oktaviana et al., 2022).

$$K(x_i, x_j) = exp\left(-\gamma \left|\left|x_i - x_j\right|\right|^2\right) \qquad (2)$$

with xi, xj is a pair of training data, and ϕ is a mapping function from inner space into feature space. Parameter γ (gamma) is a positive parameter that influences the width of the curve in the Support Vector Machine (SVM). The larger the gamma value, the narrower the curve. The classification process for a data object x can be formulated as follows: (Kavabilla et al., 2023).

$$f(\Phi(x)) = sign\left(\sum_{i=1}^{p} \alpha_i y_i K(x_i, x_j) + b\right) \qquad (3)$$

$$f(\Phi(x)) \begin{cases} 1, if \sum_{i=1}^{p} \alpha_i y_i K(x_i, x_j) + b \geq 0 \\ -1, if \sum_{i=1}^{p} \alpha_i y_i K(x_i, x_j) + b < 0 \end{cases} \qquad (4)$$

Where :

$f(\Phi(x))$	: Result data x
yi	: Data classs
αi	: Coefisien lagrange
$K(\boldsymbol{xi}, \boldsymbol{xj})$	: Function kernel test data and train data
b	: Bias
p	: More support vector

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 3, July 2024 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i3.13733 p-ISSN : 2541-044X

Initially, SVM was designed to solve binary classification tasks, but its functions have been extended to tackle multi-class classification problems through the use of pattern recognition. Pattern recognition is a technique that links data to predetermined categories. Although SVM was originally a linear classifier, it evolved to handle non-linear problems by introducing the Kernel concept in high-dimensional workspaces. In this high-dimensional workspace, SVM looks for a hyperplane that can maximize the distance between data classes (Rahman et al., 2021). In the context of this research, the SVM technique is recognized as a more sophisticated machine learning method compared to previous methods, such as Neural Networks (NN) (Utama et al., 2019).

Classification

Classification is the process of developing a model to differentiate between one class and another, intending to use the model to predict the class of objects whose class is not yet known or to study a set of data to generate rules that can classify or recognize new data that has never been studied (Halim & Purba, 2021). Text classification is a step to identify new patterns that have not been revealed before (Herianto, 2019). The purpose of classification is to assign classes to data based on learned models, and this classification can include various classes, not limited to just one. Several general stages in the classification process include pre-processing stage, weighting, and classification process (Nanda et al., 2022). Pre-processing stage is an important first step to ensure data processing can be carried out smoothly. Weighting is process of assigning weights to data, especially text data. Classification process is the step taken after all the data has been cleaned and weighted, and can be calculated using several methods. In the learning process, input is required in the form of a series of labeled training data (which has class attributes), and the result is a classification model (Halim & Purba, 2021).

Evaluation

Evaluation of a classification model can be carried out on test data that has specific values and is not used for training data. Model classification involves creating a representation of a row of data with the output being a target prediction or data class. The classification that has two class outputs is called binary classification, where both classes are described as {positive, negative}. The measurement used in evaluating the classification model is accuracy, which is the ratio of correct predictions to the total number of predictions. This evaluation aims to measure accuracy using K-fold cross-validation (Hasibuan & Serdano, 2022). The evaluation process uses a confusion matrix to see the performance of the classification model and determine its accuracy.

$$Accuracy = \frac{Number\ of\ Correct\ Prediction}{Total\ Number\ of\ Prediction} \qquad (5)$$

For binary classification, accuracy can also be calculated in positive and negative terms as follows

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (6)$$

Where :
TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

The evaluation process uses a confusion matrix. This evaluation process serves to see the performance of the classification model that has been processed and determine its accuracy. This method uses the matrix table in Table 1.

Table 1. Confusion Matrix

		Prediction		Total
		(Negative)	(Positive)	
Example	(Negative)	p	q	p+q
	(Positive)	u	v	u+v
	Total	p+u	q+v	M

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 3, July 2024 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i3.13733 p-ISSN : 2541-044X

where :

a) p is the number of accurate predictions that the instance is negative (tn).
b) q is the number of accurate predictions that the instance is positive (fn).
c) u is the number of accurate predictions that the instance is negative (fp).
d) v is the number of accurate predictions that the instance is positive (tp).

METHOD

In this research, the type of research used is Development Research or Research and Development (R&D). R&D is used to produce a particular product or model and test the product. In the framework of this research, the first step is to obtain data. The data used is a film review dataset taken from the Kaggle.com site. After the data is obtained, the next step is preprocessing. The results of the preprocessing are then divided into a training set and a testing set with an 80:20 split. The 80% training data is used to find the form of polarization, and this training lexicon is used for training data in SVM. To build this research, the author created a research framework in the form of a research flowchart, which can be seen in Figure 1. The research framework in Figure 1 first involves preprocessing the data, including cleaning, case folding, tokenizing, and stemming. After that, the data is split 80:20, with 80% of the training data labeled by the lexicon-based method, while the remaining 20% is used as test data. The process of separating the two parts of the data is done randomly. Model testing will indicate how accurate the SVM is on the data.

Figure 1 Research Framework

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 3, July 2024 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i3.13733 p-ISSN : 2541-044X

Data Collection

The data was retrieved from the Kaggle.com site in the form of IMDB film review data for 2022. The data consists of two columns: one for numerical ratings and another for reviews. The dataset comprises 40,000 rows of data. The format of the data can be seen in Table 2.

Table 2 Dataset

No	Review
1	One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked. They are right, as this is exactly what happened with me. The first thing that struck me about Oz was its brutality and unflinching scenes of violence, which set in right from the word GO. Trust me, this is not a show for the faint hearted or timid. This show pulls no punches with regards to drugs, sex or violence. Its is hardcore, in the classic use of the word. It is called OZ as that is the nickname given to the Oswald Maximum Security State Penitentary. It focuses mainly on Emerald City, an experimental section of the prison where all the cells have glass fronts and face inwards, so privacy is not high on the agenda. Em City is home to many..Aryans, Muslims, gangstas, Latinos, Christians, Italians, Irish and more....so scuffles, death stares, dodgy dealings and shady agreements are never far away. I would say the main appeal of the show is due to the fact that it goes where other shows wouldn't dare. Forget pretty pictures painted for mainstream audiences, forget charm, forget romance...OZ doesn't mess around. The first episode I ever saw struck me as so nasty it was surreal, I couldn't say I was ready for it, but as I watched more, I developed a taste for Oz, and got accustomed to the high levels of graphic violence. Not just violence, but injustice (crooked guards who'll be sold out for a nickel, inmates who'll kill on order and get away with it, well mannered, middle class inmates being turned into prison bitches due to their lack of street skills or prison experience) Watching Oz, you may become comfortable with what is uncomfortable viewing....thats if you can get in touch with your darker side.
2	A wonderful little production. The filming technique is very unassuming- very old-time-BBC fashion and gives a comforting, and sometimes discomforting, sense of realism to the entire piece. The actors are extremely well chosen- Michael Sheen not only "has got all the polari" but he has all the voices down pat too! You can truly see the seamless editing guided by the references to Williams' diary entries, not only is it well worth the watching but it is a terrificly written and performed piece. A masterful production about one of the great master's of comedy and his life. The realism really comes home with the little things: the fantasy of the guard which, rather than use the traditional 'dream' techniques remains solid then disappears. It plays on our knowledge and our senses, particularly with the scenes concerning Orton and Halliwell and the sets (particularly of their flat with Halliwell's murals decorating every surface) are terribly well done.
40000	No one expects the Star Trek movies to be high art, but the fans do expect a movie that is as good as some of the best episodes. Unfortunately, this movie had a muddled, implausible plot that just left me cringing - this is by far the worst of the nine (so far) movies. Even the chance to watch the well known characters interact in another movie can't save this movie - including the goofy scenes with Kirk, Spock and McCoy at Yosemite. I would say this movie is not worth a rental, and hardly worth watching, however for the True Fan who needs to see all the movies, renting this movie is about the only way you'll see it - even the cable channels avoid this movie.

Pre-Processing

There are five stages in pre-processing, namely cleaning, case folding, tokenizing, filtering, and stemming. The data cleaning process is carried out to remove empty data, duplicates, punctuation, symbols, and links. The data we have is not always structured and consistent in the use of capital letters. Therefore, case folding plays an important role in generalizing the use of capital letters. For example, the text data "DaTA SCIence" will be changed to "data science" with case folding, so that all letters

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 3, July 2024 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i3.13733 p-ISSN : 2541-044X

become lowercase. In the process of data analysis, we need to break down sentences into individual words, which is called tokenizing. Tokenizing allows us to distinguish between word separators and the words themselves. The filtering stage is used to select important words from the tokenizing results. Common words that appear frequently and have no special meaning, such as conjunctions ("and", "which", "and", "after", etc.), are called stopwords. Stopword removal can reduce index size, processing time, and noise levels in the data. The stemming stage aims to reduce the number of different indices in one dataset by returning words that have suffixes or prefixes to their basic form. Apart from that, stemming also helps group words that have similar base words and meanings, even though they have different forms because they have different affixes.

Split Data

Data splitting is the process of dividing data into specific proportions. In this research, the data is divided into an 80:20 ratio, where 80% is allocated for training data and 20% for test data. The breakdown of the data can be observed in Table 3.

Table 3 Split Dataset

Train Data	80	32000 data
Test Data	20	8000 data

RESULT

Pre-Processing Results

The results of the data preprocessing process are presented in Table 2. The dataset is divided into training data and test data with a proportion of 80:20. The preprocessing process is performed on the training dataset because this dataset is used to train the model. Following the preprocessing process, duplicate data was found and subsequently deleted. After the preprocessing process, the amount of filtered data was 39,722 records. The preprocessing process can be seen in Table 4.

Table 4 Pre-Processing Results

Index	Text Clean
0	i grew up (b. 1965) watching and loving the thunderbirds. all my mates at school watched. we played "thunderbirds" before school, during lunch and after school. we all wanted to be virgil or scott. no one wanted to be alan. counting down from 5 became an art form. i took my children to see the movie hoping they would get a glimpse of what i loved as a child. how bitterly disappointing. the only high point was the snappy theme tune. not that it could compare with the original score of the thunderbirds. thankfully early saturday mornings one television channel still plays reruns of the series gerry anderson and his wife created. jonatha frakes should hand in his directors chair, his version was completely hopeless. a waste of film. utter rubbish. a cgi remake may be acceptable but replacing marionettes with homo sapiens subsp. sapiens was a huge error of judgment.
1	when i put this movie in my dvd player, and sat down with a coke and some chips, i had some expectations. i was hoping that this movie would contain some of the strong-points of the first movie: awsome animation, good flowing story, excellent voice cast, funny comedy and a kick-ass soundtrack. but, to my disappointment, not any of this is to be found in atlantis: milo's return. had i read some reviews first, i might not have been so let down. the following paragraph will be directed to those who have seen the first movie, and who enjoyed it primarily for the points mentioned.when the first scene appears, your in for a shock if you just picked atlantis: milo's return from the display-case at your local videoshop (or whatever), and had the expectations i had. the music feels as a bad imitation of the first movie, and the voice cast has been replaced by a not so fitting one. (with the exception of a few characters, like the voice of sweet). the actual drawings isnt that bad, but the animation in particular is a sad sight. the storyline is also pretty weak, as its more like three episodes of schooby-doo than the single adventurous story we got the last time. but dont misunderstand, it's not very good schooby-doo episodes. i didnt laugh a single time, although

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 3, July 2024 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i3.13733 p-ISSN : 2541-044X

	i might have sniggered once or twice.to the audience who haven't seen the first movie, or don't especially care for a similar sequel, here is a fast review of this movie as a stand-alone product: if you liked schooby-doo, you might like this movie. if you didn't, you could still enjoy this movie if you have nothing else to do. and i suspect it might be a good kids movie, but i wouldn't know. it might have been better if milo's return had been a three-episode series on a cartoon channel, or on breakfast tv
39722	christopher lambert is annoying and disappointing in his portrayal as gideon. this movie could have been a classic had lambert performed as well as tom hanks in forrest gump, or dustin hoffman as raymond babbitt in rain man, or sean penn as sam dawson in i am sam.too bad because the story line is meaningful to us in life, the supporting performances by charlton heston, carroll o'connor, shirley jones, mike connors and shelley winters were excelent. 3 of 10.

Data Labeling Process

The data labeling process was conducted to categorize the dataset in Table 4 into positive, negative, and neutral categories using lexicon-based techniques with the VADER Sentiment library. This method determines sentiment classes based on the lexicon. With the VADER analysis method, sentiment analysis produces scores for several categories (negative, positive, neutral) and adds a total sentiment score known as the compound score (combined score). A negative score indicates negative sentiment, a positive score indicates positive sentiment and a score of 0 indicates neutral sentiment. The VADER analysis method is a Python programming language package from the NLTK (Natural Language Toolkit) feature and can be used in conjunction with the Liu Hu method and SentimentAnalyzer Tools for categorizing sentiment classes in Orange Data Mining. The source of the VADER lexicon is English. The calculated compound value can be positive, negative, or neutral. A positive value is obtained if the compound score is greater than 0, a negative value if the compound score is less than 0, and a neutral value if the compound score is equal to 0. The labeling results can be seen in Table 5.

Table 5 Lexicon Labeling Results

Index	Text	Compound Score
1	when i put this movie in my dvd player, and sat down with a coke and some chips, i had some expectations. i was hoping that this movie would contain some of the strong-points of the first movie: awsome animation, good flowing story, excellent voice cast……	0.9665
2	why do people who do not know what a particular time in the past was like feel the need to try to define that time for others? replace woodstock with the civil war and the apollo moon-landing with the titanic sinking and you've got as realistic a flick as this formulaic soap opera populated entirely by low-life trash. is this what kids who were too young to be allowed to go to woodstock and who failed grade school composition do? "i'll show those old meanies, i'll put out my own movie and prove that you don't have to know nuttin about your topic to still make money!" yeah, we already know that. the one thing watching this film did for me was to give me a little insight into underclass thinking…..	-0.9568
3	i grew up (b. 1965) watching and loving the thunderbirds. all my mates at school watched. we played "thunderbirds" before school, during lunch and after school. we all wanted to be virgil or scott. no one wanted to be alan. counting down from 5 became an art form. i took my children to see the movie hoping they would get a glimpse of what i loved as a child. how bitterly disappointing. the only high point was the snappy theme tune	0.651

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 3, July 2024
DOI : https://doi.org/10.33395/sinkron.v8i3.13733
e-ISSN : 2541-2019
p-ISSN : 2541-044X

39722	even though i have great interest in biblical movies, i was bored to death every minute of the movie. everything is bad. the movie is too long, the acting is most of the time a joke and the script is horrible. i did not get the point in mixing the story about abraham and noah together. so if you value your time and sanity stay away from this horror.	-0.7515

The visualization results of compound scores on all training data can be seen in Figures 2 and 3. In Figure 2, the number of negative values between -1 and -0.75 is the same, whereas the number of positive values between 0.75 is more than the highest value, which is 1. To see the total scores in detail, see Figure 3.

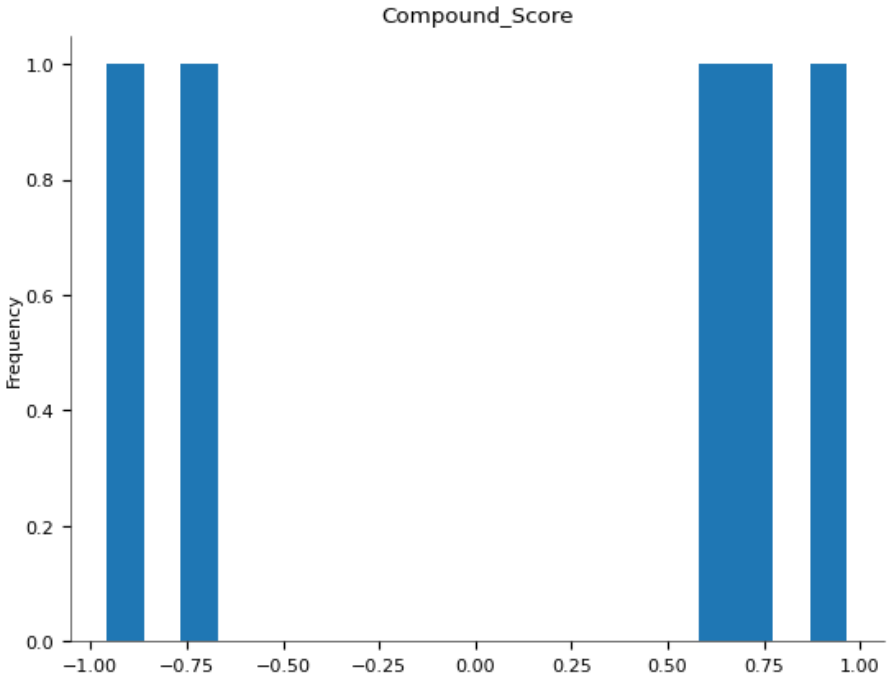

Figure 2 Score Visualization

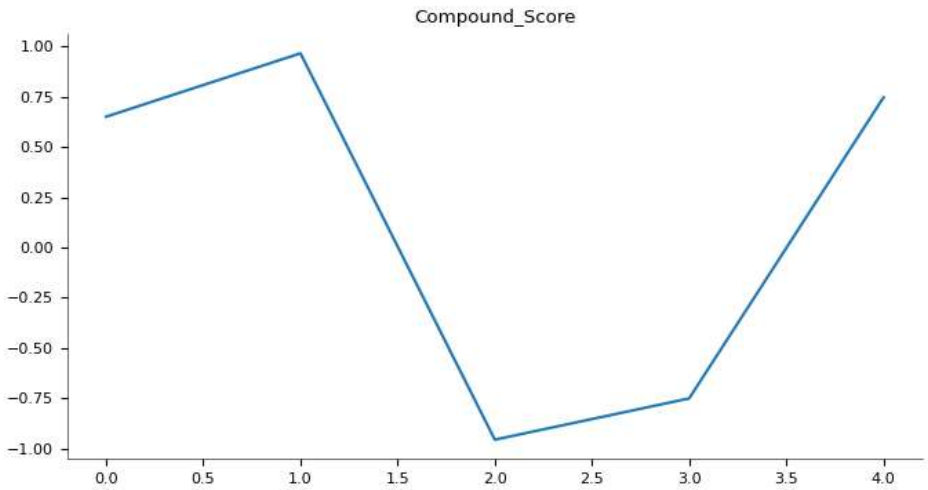

Figure 3 Movement of Compound Score Number

To make it easier to understand the score results, an equation was created to group compound scores into positive, negative, and neutral sentiment categories. This grouping can be seen in Table 6.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 3, July 2024
DOI : https://doi.org/10.33395/sinkron.v8i3.13733
e-ISSN : 2541-2019
p-ISSN : 2541-044X

```
#deskriptif commpound score
datasetclean.loc[datasetclean['Compound_Score'] < 0, 'Sentiments'] = 'Negatif'
datasetclean.loc[datasetclean['Compound_Score'] == 0, 'Sentiments'] = 'Netral'
datasetclean.loc[datasetclean['Compound_Score'] > 0, 'Sentiments'] = 'Positif'
datasetclean.head()
```

Table 6 Sentiment Labeling Results

Index	Text	*Compound* Score	Sentiment
1	i grew up (b. 1965) watching and loving the thunderbirds. all my mates at school watched. we played "thunderbirds" before school, during lunch and after school. we all wanted to be virgil or scott	0.651	Positive
2	when i put this movie in my dvd player, and sat down with a coke and some chips, i had some expectations. i was hoping that this movie would contain some of the strong-points of the first movie: awsome animation, good flowing story, excellent voice cast, funny comedy and a kick-ass soundtrack	0.9665	Positive
3	why do people who do not know what a particular time in the past was like feel the need to try to define that time for others? replace woodstock with the civil war and the apollo moon-landing with the titanic sinking and you've got as realistic a flick as this formulaic soap opera populated entirely by low-life trash. is this what kids who were too young to be allowed to go to woodstock and who failed grade school composition do? "i'll show those old meanies	-0.9568	Negative
39722	even though i have great interest in biblical movies, i was bored to death every minute of the movie. everything is bad. the movie is too long, the acting is most of the time a joke and the script is horrible. i did not get the point in mixing the story about abraham and noah together. so if you value your time and sanity stay away from this horror.	-0.7515	Negative

Visualization results of the percentage of sentiment in positive, negative, and neutral classes can be seen in Figure 4. The results show that positive sentiment has a percentage of 66.39% or around 26,372 data, negative sentiment is around 33.54% or 13,323 data, and neutral sentiment is around 0.07% or 27 data.

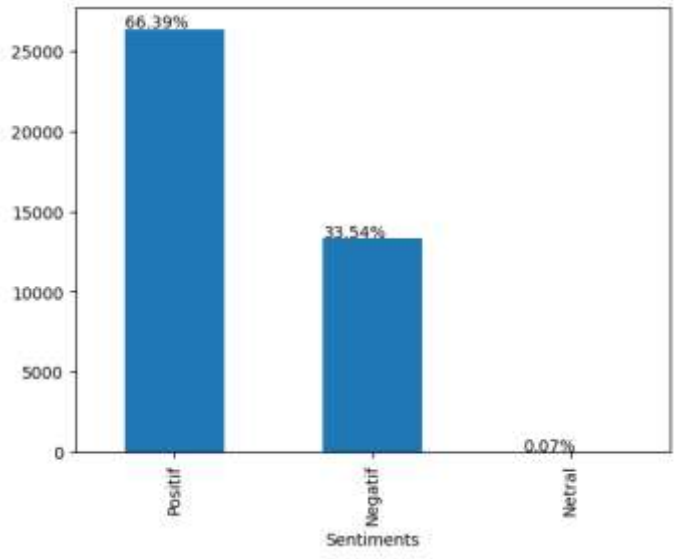

Figure 4 Number of Sentiment Percentages

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 3, July 2024 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i3.13733 p-ISSN : 2541-044X

DISCUSSIONS

SVM Modeling

The modeling process is to split the data, the data is divided into two parts, train and test data, with a percentage of 80:20. This process is carried out randomly with the library: from sklearn.model_selection import train_test_split. The results of the split data can be seen in Table 7.

Table 7 Split Data

Variable	Total
X_*Train*	31777
X_*Test*	7945
Total	39722

Test results of the Lexicon+SVM model with 6 (six) experiments with a complexity parameter value of 0.01; 0.05; 0.25; 0.5; 0.75; 1. Details can be seen in Table 8.

Table 8 Model Testing Results

Complexity to	Results
0.01	0.8563
0.05	0.8478
0.25	0.8401
0.5	0.8358
0.75	0.8328
1	0.8313

The results of the experiment above can be seen in Figure 5.

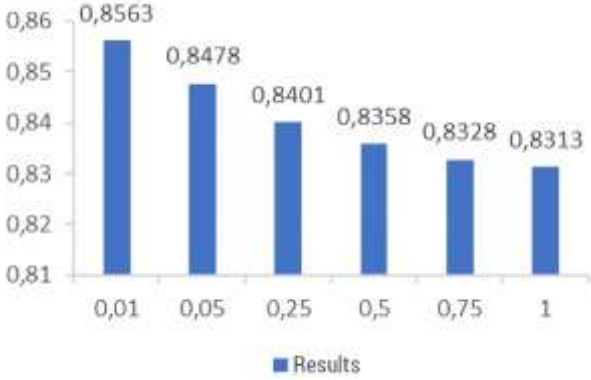

Figure 5 Results of Several Experiments

Evaluation

Based on the test results in Table 8 and Figure 5, the complexity value of 0.01 shows the highest value, namely 0.8563. This shows that this value is optimal for SVM performance in avoiding misclassification on each training data sample. The modeling carried out resulted in an overall model accuracy of around 85% based on confusion matrix calculations. The precision value, namely the positive prediction value, reached 88%, the negative prediction value 80%, and the neutral prediction value 0%. These results show that modeling using Lexicon and SVM has good performance with an accuracy result of 85%. The overall value of the confusion matrix can be seen in Table 9.

Table 9 Confusion Matrix

	precision	recall	f1-score	support
Negative	0.80	0.75	0.78	2682
Netral	0.00	0.00	0.00	1
Positive	0.88	0.91	0.89	5262
Accuracy			0.85	7945
Macro avg	0.56	0.55	0.56	7945
Weighted avg	0.85	0.85	0.85	7945

*name of corresponding author

Suggestion

It is recommended to explore techniques of text polarization other than lexicon-based ones in future research. And for further research, it is recommended to carry out comparisons with other classification algorithms, such as regression.

CONCLUSION

Based on the results of several experiments carried out in this research, several conclusions were obtained. In this research, VaderSentiment sentiment labeling from the Lexicon library is used. After preprocessing, the amount of data is reduced from 40,000 to 39,722 data. Of this data, 66.39% (around 26,372 data) is categorized as positive sentiment, 33.54% (around 13,323 data) as negative sentiment, and 0.07% (around 27 data) as neutral sentiment. Testing was carried out with six experiments using the Lexicon+SVM model. These six experiments used different complexity parameter values, namely 0.01; 0.05; 0.25; 0.5; 0.75; and 1. The best results were obtained in experiments with a complexity of 0.01, with an accuracy value of 0.8563. Based on the modeling results, the overall model accuracy is around 85%, which is calculated using the confusion matrix. The precision value, which shows the proportion of correct positive predictions, reached 88%. Meanwhile, the precision for negative predictions reaches 80%, and for neutral predictions reaches 0%. These results show that the Lexicon+SVM model has good performance, with an accuracy of 85%.

REFERENCES

Alhaq, Z., Mustopa, A., Mulyatun, S., & Santoso, J. D. (2021). Penerapan Metode Support Vector Machine Untuk Analisis Sentimen Pengguna Twitter. *Journal of Information System Management (JOISM)*, *3*(2), 44–49. https://doi.org/10.24076/joism.2021v3i2.558

Aribowo, A. S., & Khomsah, S. (2021). Implementation Of Text Mining For Emotion Detection Using The Lexicon Method (Case Study: Tweets About Covid-19). *Telematika*, *18*(1), 49. https://doi.org/10.31315/telematika.v18i1.4341

Firdaus, A., & Firdaus, W. I. (2021). Text Mining Dan Pola Algoritma Dalam Penyelesaian Masalah Informasi : (Sebuah Ulasan). *Jurnal JUPITER*, *13*(1), 66.

Furqan, M., Sriani, S., & Sari, S. M. (2022). Analisis Sentimen Menggunakan K-Nearest Neighbor Terhadap New Normal Masa Covid-19 Di Indonesia. *Techno.Com*, *21*(1), 51–60. https://doi.org/10.33633/tc.v21i1.5446

Halim, E., & Purba, R. (2021). Consumer Opinion Extraction Using Text Mining for Product Recommendations On E-Commerce. *Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)*, *4*(1), 19–28.

Hamka, M., & Ratna Sari, D. (2022). Analisis Sentimen Dan Information Extraction Pembelajaran Daring Menggunakan Pendekatan Lexicon. *Djtechno: Jurnal Teknologi Informasi*, *3*(1), 21–32. https://doi.org/10.46576/djtechno.v3i1.2194

Hasibuan, M. S., & Serdano, A. (2022). Analisis Sentimen Kebijakan Pembelajaran Tatap Muka Menggunakan Support Vector Machine dan Naive Bayes. *JRST (Jurnal Riset Sains Dan Teknologi)*, *6*(2), 199–204. https://doi.org/10.30595/jrst.v6i2.15145

Hayaty, M., & Pratama, A. H. (2023). Performance of Lexical Resource and Manual Labeling on Long Short-Term Memory Model for Text Classification. *Jurnal Ilmiah Teknik Elektro Komputer Dan Informatika*, *9*(1), 74–84. https://doi.org/10.26555/jiteki.v9i1.25375

Herianto. (2019). *Penerapan Text-Mining Untuk Mengidentifikasi*. *VIII*(2), 36–44.

*name of corresponding author

Hofmann, M., & Chisholm, A. (2016). Text Mining and Visualization: Case Studies Using Open-Source Tools. In *Text Mining and Visualization: Case Studies Using Open-Source Tools*.

Hoiriyah, H., Qomariya, N., Darmawan, A. K., Walid, M., & Efenie, Y. (2023). Sentiment Analysis on Lgbt Issues in Indonesia With Lexicon-Based and Support Vector Machine Algorithms. *Jurnal Pilar Nusa Mandiri*, *19*(1), 27–36. https://doi.org/10.33480/pilar.v19i1.4183

Jo, T. (2019). Text Mining: Concepts, Implementation, and Big Data Challenge. In *Studies in Big Data* (Vol. 45).

Kavabilla, F. E., Widiharih, T., & Warsito, B. (2023). Analisis Sentimen Pada Ulasan Aplikasi Investasi Online Ajaib Pada Google Play Menggunakan Metode Support Vector Machine Dan Maximum Entropy. *Jurnal Gaussian*, *11*(4), 542–553. https://doi.org/10.14710/j.gauss.11.4.542-553

Kusnia, U., Kurniawan, F., & Artikel, S. (2022). Analisis Sentimen Review Aplikasi Media Berita Online Pada Google Play menggunakan Metode Algoritma Support Vector Machines (SVM) Dan Naive Bayes. *Jurnal Keilmuan Dan Aplikasi Teknik Informatika*, *5*(36), 22–28.

Nanda, R., Haerani, E., Gusti, S. K., & Ramadhani, S. (2022). Klasifikasi Berita Menggunakan Metode Support Vector Machine. *Jurnal Nasional Komputasi Dan Teknologi Informasi (JNKTI)*, *5*(2), 269–278. https://doi.org/10.32672/jnkti.v5i2.4193

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

Putri, T. T. A., Mendoza, Mhd. D., & Alie, M. F. (2020). Sentiment Analysis On Twitter Using The Target-Dependent Approach And The Support Vector Machine (SVM) Method. *Jurnal Mantik*, *3*(1), 20–26.

Rahman, O. H., Abdillah, G., & Komarudin, A. (2021). Klasifikasi Ujaran Kebencian pada Media Sosial Twitter Menggunakan Support Vector Machine. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(1), 17–23. https://doi.org/10.29207/resti.v5i1.2700

Saputra, F. T., Nurhadryani, Y., Wijaya, S. H., & Defina, D. (2021). Analisis Sentimen Bahasa Indonesia pada Twitter Menggunakan Struktur Tree Berbasis Leksikon. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *8*(1), 135. https://doi.org/10.25126/jtiik.0814133

Scutelnicu, L. A. (2023). An Approach of Interconnecting Romanian Lexical Resources. *Procedia Computer Science*, *225*, 804–814. https://doi.org/10.1016/j.procs.2023.10.067

Utama, H. S., Rosiyadi, D., Prakoso, B. S., & Ariadarma, D. (2019). Analisis Sentimen Sistem Ganjil Genap di Tol Bekasi Menggunakan Algoritma Support Vector Machine. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *3*(2), 243–250. https://doi.org/10.29207/resti.v3i2.1050

*name of corresponding author