



SENTIMENT ANALYSIS OF YOUTUBE CHANNELS FOR LEARNING ENGLISH FOR TOEFL LEARNING RECOMMENDATIONS USING THE SVM METHOD

Gani Adi Alzani Rusandi^{1*}, Dede Syahrul Anwar², Rudi Hartono³

^{1,2,3}) Informatics Engineering Study Program, Faculty of Engineering, University of Struggle Tasikmalaya

Correspondence: E-mail: ganiadiar@gmail.com

ABSTRACT

The use of social media in Indonesia is not only for entertainment but also as a means of education. YouTube, as one of the most popular sites in the world, is used for learning including TOEFL exam preparation. Students of Universitas Perjuangan Tasikmalaya often have difficulty choosing the right learning resources among the many English learning channels such as Andrian Permadi , Yanto Tanjung, and Rumah Cerdas Bahasa Inggris. This study aims to overcome this problem by analyzing the sentiment of YouTube user comments on these channels using the Support Vector Machine method. The research stages include collecting comment data, data preprocessing, data labeling, and training the Support Vector Machine model for sentiment analysis. The results showed that the Yanto Tanjung channel got the highest accuracy score of 84%, making it the best choice for TOEFL preparation. The Andrian Permadi channel achieved 80% accuracy, and the Rumah Cerdas Bahasa Inggris channel achieved 75% accuracy. The contribution of this study is to provide recommendations based on sentiment analysis to help students choose the right YouTube channel for learning English, thereby maximizing their preparation for the TOEFL exam.

ARTICLE INFO

Keyword:

*Analisis Sentimen,
YouTube,
Pembelajaran Bahasa Inggris,
Support Vector Machine.*

Introduction

YouTube is one of the most visited sites by internet users worldwide. This site, which is a social media, allows sharing videos with various contents, themes and topics between users [1]. According to Statista [2], YouTube is the most widely used social media in Indonesia, with a total of 139 million users in early 2023. Facebook is in second place with 119.9 million users.

Currently, there are many content creators who discuss English learning materials as an alternative learning resource. There are several content creators who create content related to English learning, including Yanto Tanjung, Andrian Permadi, and Rumah Cerdas Bahasa Inggris. Currently, content creator Yanto Tanjung has 144,000 subscribers with a total of 7 million views. Then Andrian Permadi has 150,000 subscribers with a total of 6 million views. Meanwhile, the content creator Rumah Cerdas Bahasa Inggris has 84,300 subscribers with a total of 3 million views.

Universitas Perjuangan Tasikmalaya is a private university located in Tasikmalaya City, West Java. [3]Every year, the *TOEFL (Test of English as a Foreign Language)* test is always held. The *TOEFL test* is a mandatory graduation requirement for Universitas Perjuangan Tasikmalaya students, so it is important to choose a YouTube *channel* for teaching materials as complementary material in practicing English for TOEFL preparation. With the increasing number of YouTube *channels* that provide English learning, especially in preparing for the *TOEFL test*, students at Universitas Perjuangan Tasikmalaya face difficulties in choosing additional learning resources that are suitable and effective.

This is due to the many choices of learning *channels* on YouTube that offer various materials and learning methods. As a result, they are often confused in choosing a learning *channel* that suits their needs and interests. From the *TOEFL exam results data* for the 2019-2023 intake, 5176 participants took the exam. Of that number, 73.65% of participants scored less than 400, while 26.35% of participants scored more than 400. Students at the University of Perjuangan Tasikmalaya averaged a total exam score of 379.

Based on the results of a questionnaire conducted by researchers with 33 respondents of Informatics Engineering students at the University of Perjuangan Tasikmalaya on the question "Are you currently preparing for the *TOEFL exam*?" produced an answer of 84.8% answered YES and 15.2% answered NO. While on the question "Do you feel confused or confused when trying to determine the right YouTube channel for *TOEFL preparation*?" produced an answer of 100% answered YES. It can be concluded that on average students are still confused in determining the right YouTube channel for *TOEFL preparation*.

To overcome this problem, a sentiment analysis study was conducted through YouTube user comments to see public opinion from each English YouTube *channel*. *This study is also used to provide recommendations regarding the best YouTube channels in TOEFL learning* for students of Universitas Perjuangan Tasikmalaya.

According to the description of the problem, the researcher conducted a research analysis of public sentiment towards YouTube *channels* as teaching materials/learning media in preparing for the *TOEFL test* using the *Support Vector Machine (SVM)* method. With the hope that the

results of this study can provide recommendations regarding the best YouTube *channels* in *TOEFL learning* for students of the University of Perjuangan Tasikmalaya.

According to Pamungkas [4] with the title "Sentiment Analysis with SVM, NAIVE BAYES and KNN for the Study of Indonesian People's Responses to the Covid-19 Pandemic on Twitter Social Media" concluded that Support Vector Machine is the most suitable algorithm for classifying data on Indonesian people's responses to Covid-19 on Twitter social media compared to naive bayes and K-nearest neighbor.

Therefore, the researcher conducted a research on sentiment analysis which was submitted as a thesis report with the title "SENTIMENT ANALYSIS OF ENGLISH LEARNING YOUTUBE CHANNELS FOR TOEFL LEARNING RECOMMENDATIONS USING THE SVM METHOD".

2. Method

Identification of problems

Based on the questionnaire in this study, students of Universitas Perjuangan Tasikmalaya found that on average students are still confused in determining the right YouTube *channel* for *TOEFL preparation*. The solution offered is sentiment analysis through YouTube user comments to see public opinion from each English YouTube *channel*. It is hoped that the results of this study can provide accurate recommendations regarding the best YouTube *channels* in *TOEFL learning* for students of Universitas Perjuangan Tasikmalaya.

Data collection

The data collection process was carried out using an extension from Google Chrome, namely Instant Data Scraper. The data taken was a review of comments from YouTube viewers on the Yanto Tanjung, Andrian Permadi, Rumah Cerdas Bahasa Inggris channels. The data taken from each YouTube channel amounted to 1500, so the total data reached 4500. The data that was successfully collected was then saved in CSV format.

Pre-processing

Pre-processing technique is used in research with data collection for the purpose of changing raw data into a form that is easier to understand. The *pre-processing stages* in this study include *data cleaning*, *case folding*, tokenization, and *stopwords*.

Data Labeling

The labeling process is done automatically using *TextBlob*, *TextBlob* will calculate the polarity and subjectivity values. Polarity indicates the sentiment tendency of a text, while subjectivity shows how much the text contains opinion or fact. The higher the subjectivity of a text, the more likely it is an opinion, while high polarity indicates positive emotions. Based on its polarity value, the text is classified into one of three classes: positive, negative, or neutral [5].

Term Frequency – Inverse Document Frequency (TF-IDF)

After completing the data cleaning process and *pre-processing stage*, the next step is to

convert all data into numbers to provide a weight value so that the data can be processed by the classification algorithm . Determination of the weight value is based on the context or meaning of the words in the text. In this study, word weighting uses the TF-IDF method [6].

Split Data

Split data is done as a step to evaluate model performance. Thus, the dataset is partitioned into training data and test data.

SVM Classification

The concept of classification with *Support Vector Machine* (SVM) is to find the optimal hyperplane that can separate two classes of data, namely the positive class and the negative class [7].

Confusion Matrix Evaluation

The evaluation aims to assess the extent of the accuracy of the model used. By using the *confusion matrix table* , to analyze how good the accuracy of a *Support Vector Machine* (SVM) method is [8].

Data Visualization

In the data visualization stage, the number of positive, neutral and negative sentiment data from the entire dataset is represented in the form of a graph. The main objective of this stage is to provide an understanding of the public's response to the Andrian Permadi , Yanto Tanjung, and Rumah Cerdas channels in English [9].

Prediction

In this final stage, testing is carried out to assess whether the model used in sentiment analysis operates effectively or experiences errors [10]. The stages of the research flow can be seen in Figure 1.

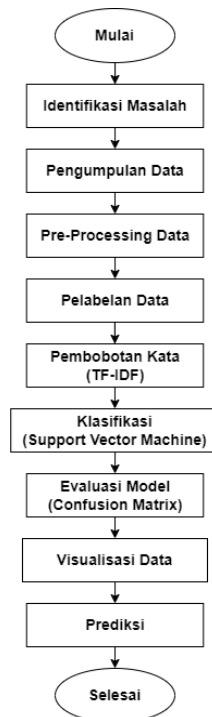


Figure 1. Research Process Flow

3. Results and Discussion

Data collection

These three channels were selected based on top searches on the YouTube application and recommendations from several questionnaire respondents. Andrian Permadi's YouTube channel has 170,000 subscribers with 8,204,326 views. Yanto Tanjung's YouTube channel has 154,000 subscribers with 7,820,783 views. Meanwhile, Rumah Cerdas Bahasa Inggris's YouTube channel has 85,000 subscribers with 3,493,729 views.

The data was collected using the Instant Data Scraper extension from Google Chrome. This extension allows users to extract information from various websites with ease. Equipped with artificial intelligence, this extension is able to identify and extract relevant data before exporting it into easy-to-process formats, such as Excel, CSV, or JSON.

Data Pre-processing

By going through a series of *pre-processing stages*, the dataset can be transformed into a cleaner, more uniform, and relevant format for further analysis. This study involves four steps in the *pre-processing process*, as follows:

1) Data Cleaning

This step aims to remove punctuation and numbers from the text, thus improving the quality of text analysis by retaining relevant words. This can be seen in the Table below.

Table 1Cleaning Channel Andrian Permadi



Text	<i>Cleaning</i>
Hopefully I can get the best 100 on my TOEFL test tomorrow  	I get the best result on my TOEFL test tomorrow.

Table 2Cleaning Channel Yanto Tanjung



Text	<i>Cleaning</i>
Thank you very much sir for the knowledge 	Thank you very much sir for the knowledge

Table 3. 3Smart Home Cleaning Channels

Text	<i>Cleaning</i>
.. waiting for the next video ...and this is very helpful.. thank you very much sir 	This video is very helpful, thank you very much, sir.

2) Case Folding

This step converts all characters in the text to lowercase to ensure consistency in text analysis, so that there is no difference between uppercase and lowercase letters during the text processing process. This can be seen in the table below.

Table 4 Casefolding Channel Andrian Permadi



Text	<i>Case Folding</i>
Hopefully I can get the best 100 on my TOEFL test tomorrow  	I get the best result on my TOEFL test tomorrow

Table 5 Casefolding Channel Yanto Tanjung



Text	Casefolding
Thank you very much sir for the knowledge 	Thank you very much sir for the knowledge

Table 6 Smart Home Channel Casefolding

Text	Casefolding
.. waiting for the next video ...and this is very helpful.. thank you very much sir 	This video is very helpful, thank you very much, sir.

3) Tokenization

This step, divides the words in a sentence, separating the text into parts of words or tokens. This is done to identify the origin of the words in the text. More information can be found in the Table below.

Table 7 Andrian Permadi Channel Tokenization



Text	Tokenization
Hopefully I can get the best 100 on my TOEFL test tomorrow  	Hopefully, tomorrow, for the TOEFL test , I can get the best result

Table 8 Tokenization of Yanto Tanjung Channel


Text	Tokenization
Thank you very much sir for the knowledge 	Thank you very much, sir, for your knowledge.

Table 9 Smart Home Channel Tokenization

Text	Tokenization
.. waiting for the next video ...and this is very helpful.. thank you very much sir 🌟	for, this video , is very, helpful, thank you , a lot, sir

4) *Stopword*

In this step, we eliminate common words that are often considered to have no important meaning or do not provide meaningful information in text analysis, such as "yang", "dan", and "dengan". More information about this process can be found in the Table below.

Table 3. 10 Stopwrod Channel Andrian Permadi

Text	<i>Stopword</i>
Hopefully I can get the best 100 on my TOEFL test tomorrow 🌟	hopefully tomorrow's TOEFL test will be the best

Table 3. 11 Stopword Channel Yanto Tanjung

Text	<i>Stopword</i>
Thank you very much sir for the knowledge 🌟	thanks for the knowledge

Table 12 Smart Home Channel Stopword

Text	<i>Stopword</i>
.. waiting for the next video ...and this is very helpful.. thank you very much sir 🌟	video helps thanks

Labeling Data

After going through the *pre-processing stage*, the data is given a label that refers to the positive, neutral and negative sentiment dictionary. The results of adding this label will provide information about the sentiment or opinion of each comment in the dataset. This information can be seen in the table below.

Table 3. 13Permadi Channel Labeling	
Text	Label
Great bro, the video is useful, thank God I passed the TOEFL in November	Positive
found the channel thank you good luck	Positive

Table 14Yanto Tanjung Channel Labeling	
Text	Label
Thank you, the material is easy to understand.	Positive
Praise be to Allah the khairan	Neutral

Table 3. 15Smart Home Channel Labeling	
Text	Label
Thanks	Positive
video helps thanks	Positive

Term Frequency – Inverse Document Frequency (TF-IDF)

TF-IDF gives higher value to words that occur frequently in a single document but rarely occur across the entire document collection. This makes the text representation more accurate in the analysis because more important words are given greater emphasis. TF-IDF weighting involves three stages, namely the calculation of Term Frequency (TF), Inverse Document Frequency (IDF) and the result of multiplying TF by IDF.

Table 3. 16Example Sentences	
Text	

Document (D1)	thank you very cool explanation
Document (D2)	crazy wrong must learn
Document (D3)	please send number
Document (D4)	crazy so hard
Document (D5)	Thank you for your answer, thank God I passed

Term Frequency Formula :

$$TF(t, d) = IDF(t, d)$$

Information:

D = total number of documents

t = term

IDF(t, d) = number of *terms* (t) in document (d)

TF(t, d) = *Term Frequency* results

Inverse Document Frequency (IDF) is a calculation of whether a word is commonly or rarely used in a document.

$$IDF = \log \frac{D}{df_t}$$

Information:

D = total number of documents

Df_t = number of words in *term* (t) in document (d)

Table 17TF Results

Terms / Say	TF						D F	IDF = $\log \left(\frac{D}{df} \right)$
	D 1	D 2	D 3	D 4	D 5	D 6		
Thank	1	0	0	0	1	1	3	0.426
You								

explana tion	1	0	0	0	0	0	1	0.903
very	1	0	0	0	0	0	1	0.903
Cool	1	0	0	0	0	0	1	0.903
gila	0	1	0	1	0	0	2	0.602
salah	0	1	0	0	0	0	1	0.903
mesti	0	1	0	0	0	0	1	0.903
belajar	0	1	0	0	0	0	1	0.903

Inverse Document Frequency (IDF) result

TF – IDF Calculation

$$TF - IDF = TF \times IDF$$

Table 3. 18TF-IDF Results

Terms / Say	TF			IDF	Weight(W) = TF*IDF		
	D 1	D 3	D 6		D1	D 3	D6
Thank You	1	0	1	0.42 6	0.42 6	0	0.4 26
explanati on	1	0	0	0.90 3	0.90 3	0	0
very	1	0	0	0.90 3	0.90 3	0	0
Cool	1	0	0	0.90 3	0.90 3	0	0
Crazy	0	0	0	0.60 2	0	0	0
Wrong	0	0	0	0.90 3	0	0	0
mesti	0	0	0	0.90 3	0	0	0

Terms / Say	TF			IDF	Weight(W) = TF*IDF		
	D 1	D 3	D 6		D1	D 3	D6
belajar	0	0	0	0.90 3	0	0	0

Split Data

Before building a sentiment analysis model, the data is divided into training data and test data with a ratio of 80:20, which means 80% is used for training and 20% is used for testing data. According to the Pareto Principle, a common and frequently used ratio is 80:20. This principle can be useful in dividing training data and testing data in many cases.

Support Vector Machine (SVM) Classification

After dividing the dataset into training data and test data, and converting text to numeric, the next step is to train the model using the training data and evaluate its performance using the test data.

SVM formula:

$$w \cdot x_i + b = 0$$

Information:

weight vector

x_i = feature vector

b = bias value

To minimize the value of SVM, you can use the formula to determine the value of b , b is the bias that must be found.

Formula to determine the value of b :

$$b = \frac{-1}{2} (\max_{i:y_i} (w \cdot x_i) + \min_{i:y_i} (w \cdot x_i))$$

With the provision of:

$$y_i(w_1 \cdot x_1 + w_2 \cdot x_2 + b) \geq 1$$

$$y_i(w_1 \cdot x_1 - w_2 \cdot x_2 - b) \leq -1$$

The value of w is the weight value of the training data, then determine the value of b using the formula to determine the value of b . After that, use w and b to classify the test data. The steps are:

1. Determine the weight vector w from the training data.
 w has been obtained from the TF-IDF weighting / calculation process, the weight vector w is written with the following conditions:

$$D1: (w_1 \cdot 0.426) + (w_2 \cdot 0.903) + (w_3 \cdot 0.903) + (w_4 \cdot 0.903) + b \geq 1$$

$$D2: (w_1 \cdot 0.602) + (w_2 \cdot 0.903) + (w_3 \cdot 0.903) + (w_4 \cdot 0.903) - b \leq -1$$

$$D3: (w_1 \cdot 0.602) + (w_2 \cdot 0.903) + (w_3 \cdot 0.903) + (w_4 \cdot 0) + b = 0$$

$$D4: (w_1 \cdot 0.602) + (w_2 \cdot 0.602) + (w_3 \cdot 0.903) + (w_4 \cdot 0) - b \leq -1$$

$$D5: (w_1 \cdot 0.426) + (w_2 \cdot 0.903) + (w_3 \cdot 0.903) + (w_4 \cdot 0.903) + b \leq 1$$

Next, find the maximum and minimum values $w \cdot x_i$ according to the class 1 or -1. Here are the steps:

For class -1 (negative):

$$\begin{aligned} \max_{i: y_i} &= -1(w \cdot x_i) &&= \max(D2, D4) \\ &&&= (0.602, 0) \\ &&&= 0.602 \end{aligned}$$

$$\begin{aligned} \min_{i: y_i} &= 1(w \cdot x_i) = \min(D1, D5) \\ &= (0.426 + 0.903 + 0.903 + 0.903, 0.426 + \\ &\quad 0.903 + 0.903 + 0.903) \\ &= \min(3.135, 3.135) \\ &= 3.135 \end{aligned}$$

2. Calculate the value of b using the formula.

$$b = \frac{-1}{2} (0.602 + 3.135) = -1.868$$

3. Use w and b for test data classification. If the value of $w \cdot x_i + b$ is positive, then the sample is classified as a positive class. Conversely, if the result is negative, then the sample will be classified as a negative class.

a. D6: "thank you the video helps"

$$\begin{aligned} w \cdot x_6 + b &= (0.426 \times 1 + 0.903 \times 1 + 0.903 \times 1 + 0) + (-1.868) = 2.232 - 1.868 \\ &= 0.364 \end{aligned}$$

Since the value is positive, D6 is classified as positive.

b. D7: "difficult to understand"

$$\begin{aligned}
 w \cdot x_7 + b &= (0.602 \times 1 + 0.903 \times 1 + 0 + 0) + \\
 &(-1.868) = 1.505 - 1.868 \\
 &= -0.363
 \end{aligned}$$

Since the value is negative, D7 is classified as negative.

In the SVM classification process using the sklearn.svm library . The module applied is SVC (*Support Vector Classifier*) for word separation shown in Figure 2 and the training process on the SVM model in Figure 3 on the Andrian Permadi , Yanto Tanjung, and Rumah Cerdas Bahasa Inggris channels .

```
from sklearn.svm import SVC
```

Figure 2. SVC (Support Vector Classifier)

```
model = SVC(kernel='linear')
model.fit(data_train, y_train)
```

SVC

SVC(kernel='linear')

Figure 2. Training Process on SVM Model

Confusion Matrix Evaluation

After the classification process with SVM is complete, the last step is to conduct an evaluation using the *Confusion Matrix* to assess performance. The results of the model evaluation are displayed in the form of a 3x3 *Confusion Matrix* and a *classification report* that includes *accuracy*, *precision*, *recall* and *F1-score*. Below are the evaluation results from the Andrian Permadi , Yanto Tanjung, and Rumah Cerdas Bahasa Inggris channels .

- a. Andrian Permadi's *channel* used 275 test data from a total of 1,371, resulting in a model evaluation reaching an accuracy level of 80%. In addition, model performance was also evaluated through *precision*, *recall*, and *F1-score* for each class. *Precision* for the negative class reached 75%, neutral reached 79%, and positive reached 81%. While *Recall* for the negative class reached 41%, neutral reached 87%, and positive reached 79%, as shown in Figure 2 .

Support Vector Machine				
[[9 6 7]				
[2 116 15]				
[1 24 95]]				
Support Vector Machine				
	precision	recall	f1-score	support
negatif	0.75	0.41	0.53	22
netral	0.79	0.87	0.83	133
positif	0.81	0.79	0.80	120
accuracy			0.80	275
macro avg	0.79	0.69	0.72	275
weighted avg	0.80	0.80	0.79	275

Figure 2. Andrian Permadi's Confusion Matrix

- b. *channel* used 266 test data from a total of 1,328, resulting in a model evaluation reaching an accuracy level of 84%. In addition, model performance was also evaluated through *precision*, *recall*, and *F1-score* for each class. *Precision* for the negative class reached 83%, neutral

reached 80%, and positive reached 87%. While *Recall* for the negative class reached 45%, neutral reached 80%, and positive reached 89%, as shown in Figure 3.

Support Vector Machine

```
[[ 5  6  0]
 [ 0 83 21]
 [ 1 15 135]]
```

Support Vector Machine

	precision	recall	f1-score	support
negatif	0.83	0.45	0.59	11
netral	0.80	0.80	0.80	104
positif	0.87	0.89	0.88	151
accuracy			0.84	266
macro avg	0.83	0.72	0.76	266
weighted avg	0.84	0.84	0.84	266

Figure 3. Yanto Tanjung's Confusion Matrix

- c. English Smart Home Channel used 289 test data from a total of 1,443, resulting in a model evaluation reaching an accuracy level of 75%. In addition, model performance was also evaluated through precision, recall, and F1-score for each class. Precision for the negative class reached 57%, neutral reached 69%, and positive reached 86%. While Recall for the negative class reached 40%, neutral reached 87%, and positive reached 70%, as shown in Figure 4.

Support Vector Machine

```
[[ 8  9  3]
 [ 2 107 14]
 [ 4 40 102]]
```

Support Vector Machine

	precision	recall	f1-score	support
negatif	0.57	0.40	0.47	20
netral	0.69	0.87	0.77	123
positif	0.86	0.70	0.77	146
accuracy			0.75	289
macro avg	0.70	0.66	0.67	289
weighted avg	0.76	0.75	0.75	289

Figure 4. Smart Home Confusion Matrix

After the confusion matrix results are obtained, visualize the confusion matrix. This visualization is a graphical representation of the table used to evaluate the performance of the classification model. On the x-axis (horizontal) the classes predicted by the model are displayed while on the y-axis (vertical) the actual classes of the test data are displayed.

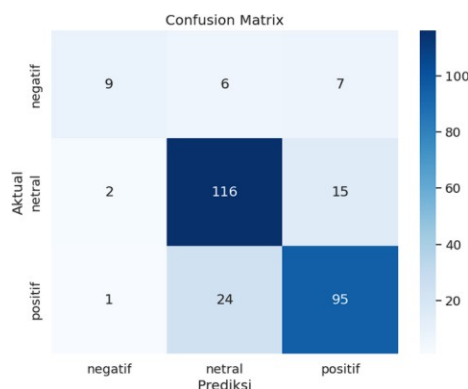


Figure 5. Visualization of Andrian Permadi's Confusion Matrix

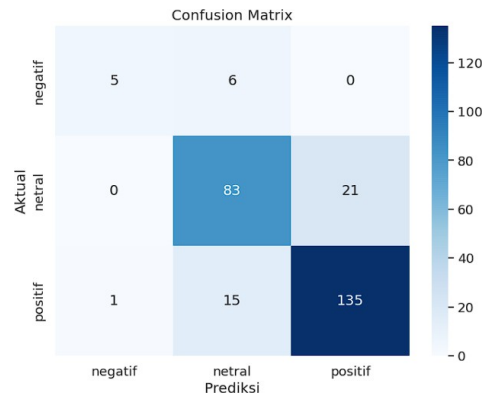


Figure 6. Visualization of Yanto Tanjung's *Confusion Matrix*

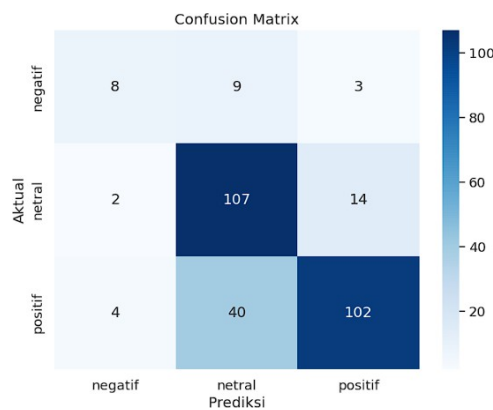
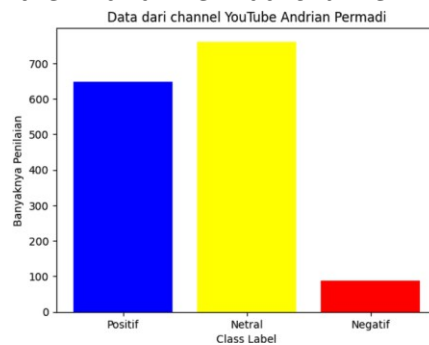


Figure 7. Visualization of Smart Home *Confusion Matrix*

Data Visualization

In this step, the positive, neutral, and negative sentiment data in the dataset as a whole are represented in the form of a graph. Thus, this visualization provides an overview of the distribution of positive, neutral, and negative data in the dataset related to *the channels of Andrian Permadi , Yanto Tanjung, and Rumah Cerdas Bahasa Inggris.*

- In Figure 8, it can be seen that there are 647 positive data, 761 neutral data, and 89 negative data from 1500 datasets on the Andrian Permadi *channel* .



Permadi's Data Visualization Results

- b. In Figure 9, it can be seen that there are 763 positive data, 724 neutral data, and 10 negative data from 1500 datasets on the Yanto Tanjung *channel* .

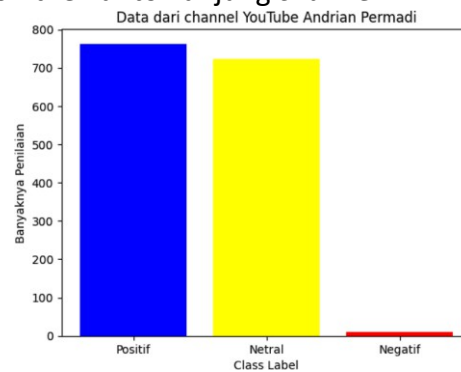


Figure 9. Yanto Tanjung's Data Visualization Results

- c. In Figure 10, it can be seen that there are 686 positive data, 713 neutral data, and 100 negative data from 1500 datasets on the English Smart Home *channel* .

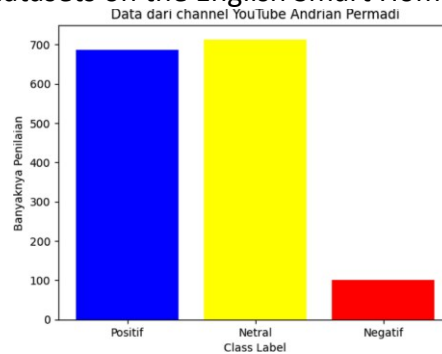


Figure 10. Smart Home Data Visualization Results

Prediction

In this step, the model that has been created is tested by predicting comments entered by the user. Each word in the comment is weighted based on the TF-IDF method. In this way, the system can determine whether the comment is classified as a positive, neutral, or negative class. For example, the comment "Thank you sir, I studied on this channel for a month , thank God I got a score of 497 on the TOEFL test on campus" will be converted into a numeric representation using the TF-IDF method, then it will be classified as a positive, neutral, or negative class based on the SVM formula. At this stage, the experiment was carried out 3 times for each *channel* , and the results were in accordance with the tone expression of the sentence entered.

4. Conclusion

1. The use of the SVM method in sentiment analysis in this study showed good performance in predicting sentiment from the data used. By using a division of training data of 80% and test data of 20%, it produces various accuracies from each *channel*, including:
 - a. *channel* achieved the highest accuracy of 84%, with a total of 763 positive sentiments, 724 neutral sentiments, and 10 negative sentiments.
 - b. Andrian Permadi's *channel* got an accuracy of 80%, with a total of 647 positive sentiments, 761 neutral sentiments, and 89 negative sentiments.

c. *Channel* , the accuracy achieved was 75%, with a total of 686 positive sentiments, 713 neutral sentiments, and 100 negative sentiments.

2. On Yanto Tanjung's *channel* , it was found that the accuracy is very high, reaching 84%. This indicates that Yanto Tanjung is the best choice as a source for learning English, especially for *TOEFL exam preparation* .

Meanwhile, for Andrian Permadi's *channel* , although the accuracy is still quite high, which is 80%, the number of negative comments received (89) is quite significant. This shows that the user experience can be improved to achieve more positive sentiment and improve user perception of this *channel* as a source of learning English for *TOEFL* .

References

- [1] R. Mtsn , L. Utara, and K. Utara, "USE OF YOUTUBE AS ENGLISH LEARNING MEDIA DURING THE COVID 19 PANDEMIC," *August* , vol. 1, no. 2, 2021.
- [2] Agnes Z. Yonatan AGNES Z. YONATAN, "A Look at Indonesian Social Media Users 2017-2026," data.goodstats.id.
- [3] Wikipedia, "Tasikmalaya University of Struggle," https://id.wikipedia.org/wiki/Universitas_Perjuangan_Tasikmalaya#cite_note-1.
- [4] FS Pamungkas and I. Kharisudin , "Sentiment Analysis with SVM," vol. 4, pp. 628–634, 2021, [Online]. Available: <https://journal.unnes.ac.id/sju/index.php/prisma/>
- [5] A. Baita and N. Cahyono, "SENTIMENT ANALYSIS ON THE SINOVAC VACCINE USING THE SUPPORT VECTOR MACHINE (SVM) AND K-NEAREST NEIGHBOR (KNN) ALGORITHMS."
- [6] D. Septiani and I. Isabela, "SINTESIA: Indonesian Journal of Information Systems and Technology TERM FREQUENCY INVERSE DOCUMENT FREQUENCY (TF-IDF) ANALYSIS IN INFORMATION RETRIEVAL IN TEXT DOCUMENTS".
- [7] Muhammad Harris Syafa'at , ER Setyaningsih , and Y. Kristian, "SVM FOR SENTIMENT ANALYSIS OF REGIONAL HEAD CANDIDATES BASED ON VIDEO COMMENT DATA ON REGIONAL ELECTION DEBATE ON YOUTUBE," *Antivirus : Scientific Journal of Informatics Engineering* , vol. 15, no. 2, pp. 262–276, Dec. 2021, doi: 10.35457/antivirus.v15i1.1539.
- [8] V. Fitriyana *et al.* , "Sentiment Analysis of Jamsostek Mobile Application Reviews Using Support Vector Machine Method," 2023.
- [9] Fauzi, A., Fa'rifah , RY, & Alam, EN (2023). Sentiment Analysis of Food and Beverage Trends with Support Vector Machine as a Recommendation of Business Opportunities for MSMEs. *Kesatria: Journal of Information System Application (Computers and Management)*, 4(4), 988-995.
- [10] Maulaya, A. K. (2022). Analisis Sentimen Menggunakan Support Vector Machine Masyarakat Indonesia Di Twitter Terkait Bjorka. *Jurnal CoSciTech (Computer Science and Information Technology)*, 3(3), 495-500.