Sentiment Analysis of Positive and Negative of YouTube Comments Using Naïve Bayes – Support Vector Machine (NBSVM) Classifier

1st Abbi Nizar Muhammad
Faculty of Computer Science
University of Jember
Jember, Indonesia
abbinizarm@gmail.com

2nd Saiful Bukhori
Faculty of Computer Science
University of Jember
Jember, Indonesia
saiful.ilkom@unej.ac.id

3rd Priza Pandunata
Faculty of Computer Science
University of Jember
Jember, Indonesia
priza@unej.ac.id

Abstract **— Sentiment analysis on the YouTube video comments is a process of understanding, extracting, and processing textual data automatically to obtain sentiment information contained in one sentence of YouTube video comment. Text mining approach becomes the best alternative to interpret the meaning of each comment. The classification of positive and negative content becomes very important for the YouTube user to assess how meaningful the content that has been published is based on user opinion. Naïve Bayes and Support Vector Machine is extensively used as a basic line in tasks related to texts but the performance varies significantly in all variants, features, and numbers of data collection. Naïve Bayes is very good in classifying texts with the small number of data and document snippets while Support Vector is very good in classifying texts with relatively many numbers of data or full-length document. The combination of Naïve Bayes and Support Vector Machine produces better accuracy level and stronger performance with the use of a 7:3 scale of data that is 70% training data and 30% testing data. By producing the highest performance test values, namely precision of 91%, recall of 83% and flscore of 87%.**

Keywords—text mining, sentiment analysis, youtube, nbsvm, natural language processing, naïve bayes, support vector machine

I. Introduction

In 2018 YouTube users reached 1.5 billion from around the world and this number is projected to grow to 1.86 billion users in 2021 [1]. YouTube is the largest video platform in the world by displaying a variety of media content created by companies or individuals that include music videos, product promotion videos, blog videos, review videos, educational videos [2]. The increase in users is directly proportional to the increasing number of video content uploaded on YouTube. This has become a place for everyone to compete in creating video content and earning income from uploaded videos.

Various reactions of opinion in the user comments when looking at the video content uploaded on YouTube are very affecting to the reputation of the video content and channel. The most important thing of information collecting is figuring out what other people do of thinking. [3]. Understanding something people think becomes the most important information for the content creator to make the acceptable video for the user.

Therefore an approach can be made to find out the perception of YouTube user on video content by using sentiment analysis data obtained from the textual content. Text mining approach becomes the best alternative to interpret the meaning of each comment. The classification of positive and negative content becomes very important for the YouTube user to assess how meaningful the content that has been published is based on user opinion comments.

This research would like to give a contribution to the development of sentiment analysis approach using Naïve Bayes - Support Vector Machine (NBSVM) method with a Binary Classification approach. The use of this method was chosen because this classification method worked very well on the document snippets and longer document to analyze the sentiment, topic and subjective classification, and frequently better than the result previously published. [4]. Naïve Bayes and Support Vector Machine is extensively used as a basic line in tasks related to texts but the performance varies significantly in all variants, features, and numbers of data collection. Naïve Bayes is very good in classifying texts with the small number of data or document snippets while Support Vector is very good in classifying texts with relatively many numbers of data or full-length document. The combination of Naïve Bayes and Support Vector Machine produces better accuracy level and stronger performance. However, the combination of several classifiers does not always increase the accuracy of classifications compared to classifications that only use one method [5].

II. Materials and Methods

A. Relate Research

Research by Ferly Gunawan, M Ali Fauzi and Putra Pandu Adikara entitled "Sentiment analysis on mobile application reviews using Naïve Bayes and Levenshtein Distance-based word normalization (Case study of BCA mobile applications)". The results of the Naive Bayes test

978-1-7281-3436-9/19/$31.00 ©2019 IEEE

without Levenshtein Distance have an accuracy value of 94.4% [6].

The research of Fatimah Wulandari and Anto Satriyo Nugroho entitled Text Classification Using Support Vector Machine for Web mining based on spatio temporal analysis of the spread of tropical diseases resulted in Support Vector Machine producing an accuracy value of 92.5% compared to Naïve Bayes Classifier whose accuracy was 90% [7].

Based on previous research compared to the Naïve Bayes and Support Vector Machine method, outperformed and worked well. The difference in results and accuracy of each study depends on the data tested how many variants, features and number of data sets were tested. Sida Wang's research entitled Baselines and Bigrams: Simple, Good Sentiment and Topic Classification illustrates the merging of the two Naïve Bayes methods and the Support Vector Machine to find higher accuracy values. The results with the film review test data resulted in an accuracy value of 91.22 [4].

B. *Sentiment Analysis*

Sentiment analysis or opinion mining is a process of understanding, extracting, and processing textual data automatically to obtain sentiment information contained in one sentence of opinion. Sentiment analysis is done to see the opinion or opinion tendency on a problem or object by someone, whether tending to have a negative or positive view or opinion [8].

C. *Text Mining*

Text mining is a pattern extraction process (information and knowledge) from a large number of unstructured data sources. Text mining has a goal and uses the same process with the data mining, however, it has a different input. The inputs for text mining are an unstructured word, such as document, word, pdf, quotes, comments, review, etc., while the input for data mining is structured data [9].

D. *Natural Language Processing*

Natural Language Processing is a part of computer science and artificial intelligence that handles human languages. The most important stage of doing text mining using natural language processing is the preprocessing stage because in general the text that will be carried out in the process of text mining has different characteristics which have a high dimension, there is noise in the data, and there is a text structure that is not good [8]. Generally, the preprocessing stage in text mining on the document is case folding, tokenization, filtering, and stemming.

E. *Classification Method*

Naïve Bayes works very well in a short sentiment task while Support Vector Machine is very good for a longer document. Hybrid Method Naïve Bayes with Support Vector Machine produces higher accuracy value with variances of Support Vector Machine which uses calculation ratio of log Naïve Bayes as a feature value consistently performing well in every task and data collection

As a linear classification model, the main model variant to predict for k testing case is as follows:

$$y^{(k)} = sign\ (w^T x^{(k)} + b) \qquad (1)$$

Thus, F^((i)) becomes a vector feature for the training case of i with the binary label of y^((i)) ∈ -1,1. Determine 2 calculations of P and Q vectors as follow

$$P = a + \sum_{i:y^{(i)}=1} f(i) \qquad (2)$$

$$q = a + \sum_{i:y^{(i)}=-1} f(i) \qquad (3)$$

Where α is a smoothing parameter and Log-count ratio can be defined as follows:

$$r = \log\left(\frac{P/\| p \| 1}{q/\| p \| 1}\right) \qquad (4)$$

1) *Multinomial Naïve Bayes*

In Naïve Bayes vector feature represents the frequency with events produced by multinomial distribution of P = (p1, …. , pn). Vector feature of x = (x1, …, x2) is histogram, where x_k is the number of k observed from several events in a certain example. With multinomial assumption and Naïve Bayes assumption, x probably has terms on y influenced by

$$P(y|x) = \log p(y) + \sum_k x_k \log p_{yk} \qquad (5)$$

Multinomial Naïve Bayes Classifier becomes linear classifier when expressed in log-space, therefore, the equation of 4 can be replaced to the equation of 1 becoming

$$y = sign\ \left(\log\frac{N_+}{N_-} + r^T f\right) \qquad (6)$$

2) *Support Vector Machine*

Support Vector Machine is a machine learning method working under the structural risk minimization (SRM) principle with the purpose to find the best hyperplane separating two classes on input space. x^((k)) = ^((k)),and w, b are obtained by minimalizing loss function

$$L(w, b) = w^T w + C \sum \max\left(0{,}1 - y^{(i)} \left(w^T F^{(i)} + b\right)\right)^2 \quad (7)$$

3) *Support Vector Machine with Naïve Bayes feature (NBSVM)*

Multinomial Naïve Bayes and Support Vector Machine works very well for all documents and the use of Naïve Bayes – Support Vector Machine (NBSVM) method uses the following model:

$$w^t = (1 - \beta)\ \omega + \beta w \qquad (8)$$

Notes:

ω = || w || 1/ | V | is the average magnitude of w

β = [0,1] is interpolation parameter

F. Data Collection Results

Data obtained to do the experiment can be obtained directly from the comments in the YouTube video content using the crawling technique with YouTube API from the writer. In extracting data from the crawling test, it uses Natural Language Processing approach to obtain accurate data.

The population from this research is YouTube video comments with education category. Data obtained to do the research is obtained directly from the comments in YouTube video content using the crawling technique with YouTube API. Channel chosen has been found based on the education category that is "*Kok Bisa ?*" channel that has a total grade of B Channel YouTube in Indonesia with the total reaching 1,416,935 subscribers and views of 142,957,953 [11].

G. Model Testing

Model testing is the stage where the researcher tests the classification model of Naïve Bayes - Support Vector Machine that has been built. Testing is done by making changes to the data scale so that it can provide maximum results. This research uses 3 different scales of data scales. Each of these scenarios is: 60:40 data scenario, 70:30 data scenario and 80:20 data scenario. The next model testing will compare the accuracy of each model using the confusion matrix which will be chosen by the model with the highest accuracy.

H. Design System Method

The problem-solving framework from the sentiment analysis process using classification method of Naïve Bayes - Support Vector Machine (NBSVM) from the system built is explained in the form of system algorithm that can be seen in the figure as follows:

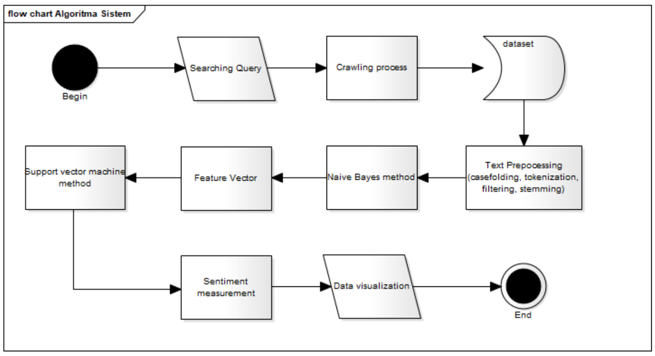
flow chart Algoritma Sistem
Begin
Searching Query
Crawling process
dataset
Text Prepocessing (casefolding, tokenization, filtering, stemming)
Naive Bayes method
Feature Vector
Support vector machine method
Sentiment measurement
Data visualization
End

Figure 1. System Algorithm

Flowchart of Classification describes the merging process of the method on training model using Naïve Bayes Support Vector Machine method.

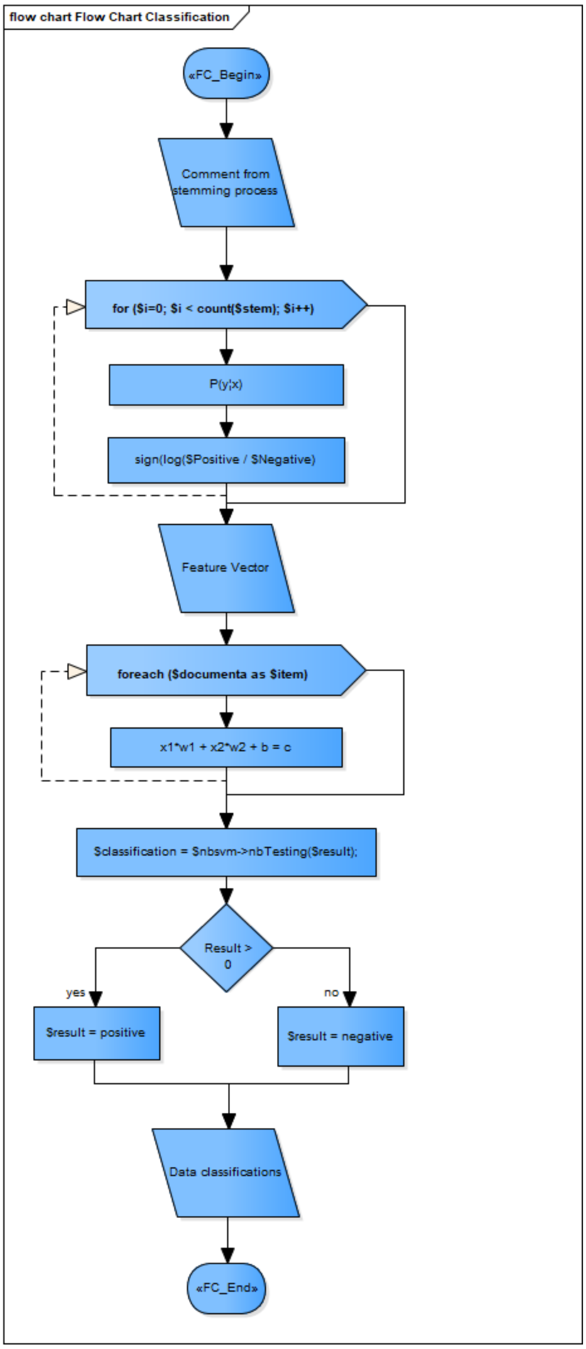
flow chart Flow Chart Classification
«FC_Begin»
Comment from stemming process
for ($i=0; $i < count($stem); $i++)
P(y|x)
sign(log($Positive / $Negative)
Feature Vector
foreach ($documenta as $item)
x1*w1 + x2*w2 + b = c
$classification = $nbsvm->nbTesting($result);
Result > 0
yes
no
$result = positive
$result = negative
Data classifications
«FC_End»

Figure 2. Flowchart of Hybrid method

III. Results and Discussion

A. Dataset Results

The type of data used is Indonesian text obtained from several sources. On the preprocessing stage on the use of stop word data in Indonesian on the stemming process uses sources from Tala with the number of 758 data [12]. While for the root word uses data from the big dictionary of Indonesian that has been existed in literature library with the number of 28,533 data and for the positive and negative word references uses data sources from the research of Wahid with the number of 3,587 data [13]. Ratio division of data use can be divided into 70% for training data and 30% for testing data. Training data used in this research amounts to 233 comments with two labels of positive and negative. While data training used amounts to 100 comments for each video. So the total data on this research amounts to 333 data.

B. Text Preprocessing Results

On the preprocessing stage, the dataset is obtained by using a repetition until all datasets are finished in the analysis with 4 stages that are casefolding, tokenization, filtering and stemming.

TABLE I. RESULTS OF TEXT PREPROCESSING

Step	Result
Youtube comment	*kontenku juga edukasi dengan animasi, walaupun masih sangat pemula... silahkan mampir bila berkenan, terima saran dan kritikan :)*
Casefolding Results	*kontenku juga edukasi dengan animasi walaupun masih sangat pemula silahkan mampir bila berkenan terima saran dan kritikan*
Tokenization Results	*kontenku juga edukasi dengan animasi walaupun masih sangat pemula silahkan mampir bila berkenan terima saran dan kritikan*
Filtering Results	*kontenku edukasi animasi pemula silahkan mampir berkenan terima saran kritikan*
Stemming Results	*konten edukasi animasi mula silah mampir kenan terima saran kritik*

C. Method Implementation Results

The implementation of Naïve Bayes Support Vector Machine method is a process to divide classes into positive and negative classes. On the Naïve Bayes method is used to calculate the possibility of occurrence of a word from changing the word into the feature vector which is later used to calculate the calculation of Support Vector Machine

1) Multinomial Naïve Bayes

The first step is making a classification model that is calculating data training by using Multinomial Naïve Bayes. This calculation sample is data training taken from 233 numbers of data training.

TABLE II. RESULTS OF NAÏVE BAYES CALCULATION

Notes	Calculation	Log space
Probability konten | Positive	$P(konten\|positive) = \frac{12+1}{1045} = 0.01244019138756$	$sign\left(\log\frac{0.01244019138756}{0.022701475595914}\right) = -1$
Probability konten | Negative	$P(konten\|negative) = \frac{18+1}{881} = 0.022701475595914$	
Probability edukasi | Positive	$P(edukasi\|positive) = \frac{11+1}{1045} = 0.011483253588517$	$sign\left(\log\frac{0.011483253588517}{0.011350737797957}\right) = 1$
Probability edukasi | Negative	$P(edukasi\|negative) = \frac{9+1}{881} = 0.011350737797957$	
Probability animasi | Positive	$P(animasi\|positive) = \frac{2+1}{1045} = 0.0028708133971292$	$sign\left(\log\frac{0.0028708133971292}{0.0011350737797957}\right) = 1$
Probability animasi | Negative	$P(animasi\|negative) = \frac{0+1}{881} = 0.0011350737797957$	
Probability mula | Positive	$P(mula\|positive) = \frac{0+1}{1045} = 0.00095693779904306$	$sign\left(\log\frac{0.00095693779904306}{0.00113507378}\right) = -1$
Probability mula | Negative	$P(mula\|negative) = \frac{0+1}{881} = 0.00113507378$	
Probability silah | Positive	$P(silah\|positive) = \frac{0+1}{1045} = 0.00095693779904306$	$sign\left(\log\frac{0.00095693779904306}{0.00113507378}\right) = -1$
Probability silah | Negative	$P(silah\|negative) = \frac{0+1}{881} = 0.00113507378$	
Probability mampir | Positive	$P(mampir\|positive) = \frac{1+1}{1045} = 0.0019138755980861$	$sign\left(\log\frac{0.0019138755980861}{0.0011350737797957}\right) = 1$
Probability mampir | Negative	$P(mampir\|negative) = \frac{0+1}{881} = 0.0011350737797957$	
Probability kenan | Positive	$P(kenan\|positive) = \frac{0+1}{1045} = 0.00095693779904306$	$sign\left(\log\frac{0.00095693779904}{0.00113507378}\right. = -1$

Probability kenan | Negative	$P(kenan\|negative) = \frac{0+1}{881} = 0.00113507378$	
Probability terima | Positive	$P(terima\|positive) = \frac{4+1}{1045} = 0.0047846889952153$	$sign\left(\log\frac{0.0047846889952153}{0.0022701475595914}\right) = 1$
Probability terima | Negative	$P(terima\|negative) = \frac{1+1}{881} = 0.0022701475595914$	
Probability saran | Positive	$P(saran\|positive) = \frac{1+1}{1045} = 0.0019138755980861$	$sign\left(\log\frac{0.0019138755980861}{0.0011350737797957}\right) = 1$
Probability saran | Negative	$P(saran\|negative) = \frac{0+1}{881} = 0.0011350737797957$	
Probability kritik | Positive	$P(kritik\|positive) = \frac{0+1}{1045} = 0.00095693779904306$	$sign\left(\log\frac{0.00095693779904306}{0.00113507378}\right) = -1$
Probability kritik | Negative	$P(kritik\|negative) = \frac{0+1}{881} = 0.00113507378$	

From the calculation of each word with positive and negative value then it produces a feature vector used to calculate the best hyperplane on the Support Vector Machine method.

2) Support Vector Machine

The next step is that the feature vector is processed to find the value of w1, w2, and b so that it gets the equation to find the best hyperplane from data training.

TABLE III. CALCULATION RESULTS OF SUPPORT VECTOR MACHINE

Notes	Calculation
Retrieving samples to find the equation of W1, W2 and b	$-0.00095693779904306W1 + -0.0034052213393871W2 + -1 = 1$ $0.0047846889952153W1 + 0.0022701475595914W2 + 1 = 1$ $0.0038277511961722W1 + -0.0011350737797957W2 + = 2$
Equation of Hyperplane	$X1 - 88.869565217391 + X2\ 110.125 + 1.0450852922821 = 0$
Information	W1 = -88.869565217391 , W2 = 110.125 and b = 1.0450852922821

Model testing is done with three different data scales with the same test data producing quite diverse values. The classification model shows optimal results and achieves the highest score on a 70% scale of training data with 30% testing data.

TABLE IV. RESULTS OF TEST DATA EXPERIMENT

Scales data	Sentiment Results		Prediction of False	Prediction of True
	Positive	Negative		
6:4	73%	27%	21	79
7:3	74%	26%	20	80
8:2	60%	40%	31	69

3) Sentiment Measurement

After determining the probability using Multinomial Naïve Bayes and finding the hyperplane using Support Vector Machine the next is testing the model with the YouTube comment data that have been passed preprocessing stage.

TABLE V. RESULTS OF SENTIMENT MEASUREMENT

No	Comment	Calculation	Sentiment
1	*konten* *edukasi* *animasi* *mula silah* *mampir* *kenan* *terima* *saran* *kritik*	- 0.0069408793149475 * -88.869565217391 + -0.010035416250554 * 110.125 + 1.0450852922821 = 0.55676800463559 0.00640696552149 * -88.869565217391 + 0.0050177081252771 * 110.125 + 1.0450852922821 = 1.0282761593206 0.0016017413803725 * -88.869565217391 + 0.00050177081252771 * 110.125 + 1.0450852922821 = 0.95799674294729 0.001067827586915 * -88.869565217391 + 0.00050177081252771 * 110.125 + 1.0450852922821 = 1.0054454296354 0.0026695689672875 * -88.869565217391 + 0.0010035416250554 * 110.125 + 1.0450852922821 = 0.91835688030063	Positive

		0.001067827586915 * -88.869565217391 + 0.00050177081252771 * 110.125 + 1.0450852922821 = 1.0054454296354	

D. Performance Test

The performance test is done to find out how accurate and precise the classification process is done by looking at the recall, precision and f1 score. Data model testing was carried out in a data scale of 8:2, 7:3, 6:4 in hopes of finding high accuracy values with the data tested.

TABLE VI. RESULTS OF PERFORMANCE TEST CALCULATION

	True Positive	True Negative	False Positive	False Negative
Predict 8:2	67	13	6	14
Predict 7:3	68	13	6	13
Predict 6:4	54	13	6	27

Precision is the number of prediction accuracy of a class on the total number of prediction classified in the class. The formula to find the precision can be written as follows [14].

$$precision\ (p) = \frac{TP}{TP + FP} \qquad (9)$$

Recall is the number ration of prediction accuracy of a class on the total number of facts classified in the class. The formula to find recall can be written as follows

$$recall\ (r) = \frac{TP}{TP + FN} \qquad (10)$$

To combine the two then used the f1 score calculation. The formula to find f1 score can be written as follows

$$f1\ score = \frac{2 * precision * recall}{precision + recall} \qquad (11)$$

From the testing results above can be concluded that the value of prediction fault is very minimum compared than the prediction result that has run well in accordance with the classification model. The testing of each experiment was also done to ensure the accuracy level obtained on each experiment was not far different from other experiments:

TABLE VII. PERFORMANCE TEST RESULTS OF EACH EXPERIMENT

	Precision	Recall	F1 Score
Scale 8:2	0.917808	0.827160494	0.87013
Scale 7:3	0.918919	0.839506173	0.877419
Scale 6:4	0.9	0.666666667	0.765957

Thus, the performance test results of positive and negative sentiments using Naïve Bayes Support Vector Machine method produces the highest precision, recall and F1 scores on the 7:3 data scale, which is 70% training data and 30% testing data.

TABLE VIII. PERFORMANCE TEST RESULTS OF CONFUSION MATRIX

Performance Test	Score
Precision	91 %
Recall	83 %
Fl score	87 %

IV. CONCLUSION

Based on the analysis and testing done, the use of text mining concept with the stages of casefolding, tokenization, filtering and stemming gave optimal result in interpreting a word making the classification model more accurate. The merging method process is started as data training is processed. Multinomial Naïve Bayes with log count ratio functions to change data text into a feature vector that will be processed within Support Vector Machine. Multinomial Naïve Bayes also has a task to determine the probability of occurrence of positive and negative data on data training. The next, the feature vector is processed into the equation of Support Vector Machine where to find the best hyperplane has to look for the distance between support vector and hyperplane so that the equation of the values W1, W2 and b found in this research is $X1 - 88.869565217391 + X2110.125 + 1.0450852922821 = 0$. In the comments that have been passing text preprocessing, the next is calculated the probability using Multinomial Naïve Bayes so that it produces a value of feature vector which is then entered into the model equation that has been decided based on data training. The classification results of Naïve Bayes – Support Vector Machine method implementation become very optimal when data training used has varied data number. On the data obtained from YouTube video comments, the combination of Naïve Bayes and Support Vector Machine methods produces better accuracy level and stronger performance with the use of 7:3 scale data, namely 70% training data and 30% testing data. By producing the highest performance test values, namely precision of 91%, recall of 83% and f1 score of 87%.

REFERENCES

[1] J. Clement, "Statista," 13 February 2018. [Online]. Available: https://www.statista.com/statistics/805656/number-youtube-viewers-worldwide/.

[2] K. Gordon, "YouTube: Statistics & Data," 2018. [Online]. Available: https://www.statista.com/topics/2019/youtube/.

[3] B. Pang dan L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval,* pp. 1-135, 2008.

[4] d. C. D. M. Sida Wang, "Baselines and Bigrams: Simple, Good Sentiment and Topic Classification," *Department of Computer Science Stanford University,* 2012.

[5] Y. H. L. A. A. K. JAIN, “Classification of Text Documents,” *THE COMPUTER JOURNAL,* pp. Vol. 41, No. 8, 1998.

[6] F. Gunawan, M. A. Fauzi dan P. P. Adikara, “Sentiment analysis on mobile application reviews using Naïve Bayes and Levenshtein Distance-based word normalization (Case study of BCA mobile applications),” *SYSTEMIC,* pp. 1-6, 2017.

[7] F. Wulandari dan A. S. Nugroho, “Text Classification Using Support Vector Machine for Webmining based on spatio temporal analysis of the spread of tropical diseases,” 2009.

[8] B. Pang, “Thumbs up? Sentiment Classification using Machine Learning,” *Association for Computational Linguistics,* pp. 79-86, 2002.

[9] R. Feldman, Advanced Approaches in Analyzing Unstructured Data, United States of America: Cambridge University Press, 2007.

[10] E. K. Steven Bird, Natural Languange Processing in Phyton, United states of america: O'reilly media, 2009.

[11] socialblade, “Socialblade,” 1 July 2019. [Online]. Available: https://socialblade.com/youtube/channel/UCu0yQD7NFMyLu_-TmKa4Hqg.

[12] I. R. Ponilan, “Pengukuran Happiness Index Masyarakat Kota Bandung pada Media Sosial,” *Ind. Symposium on Computing,* pp. 17-22, 2016.

[13] F. Z. Tala, “A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia.,” *M.Sc. Thesis. Master of Logic Project. Institute for Logic, Language and Computation. ,* 2003.

[14] D. H. &. A. S. N. Wahid, “Peringkasan Sentimen Esktraktif di Twitter Menggunakan Hybrid TF-IDF dan Cosine Similarity,” *IJCCS (Indonesian Journal of Computing and Cybernetics Systems),* pp. 10(2), 207-218, 2016.

[15] Y. Wibisono, “Klasifikasi berita bahasa indonesia menggunakan Naive Bayes Classifier,” 2005.