P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v4i1.262

ANALYSIS CLASSIFICATION SENTIMENT OF THE LARGE PRIEST OF FPI'S RETURN USING SVM CLASSIFICATION WITH OVERSAMPLING METHOD

Zetta Nillawati Reyka Putri[1], Muhammad Muhajir[2*]

[1,2]Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Islam Indonesia
17611090@students.uii.ac.id[1], mmuhajir@uii.ac.id[2*]

(*) Corresponding Author

Abstrak

Akhir tahun 2020 kepulangan Habib Rizieq ke Indonesia menuai kecaman dari masyarakat karena menyebabkan kerumunan di masa pandemi Covid-19. Berita dan opini mengenai Habib Rizieq memenuhi platform internet tak terkecuali twitter. Peneliti ingin mengklasifikasikan data teks opini kepulangan Habib Rizieq dari twitter ke sentimen positif dan negatif dengan metode Support Vector Machine. Data opini berasal dari twitter, sehingga data tersebut di analisis dengan text mining melalui tahap preprocessing. Klasifikasi SVM data tidak seimbang antara kelas positif dan negatif menghasilkan akurasi 95.06% dengan nilai presisi kelas negatif 84% dan lebih baik dari recall sebesar 72%, pada kelas positif diketahui nilai presisi sebesar 96% lebih kecil 2% dari recall 98%. Sedangkan klasifikasi svm dengan metode oversampling mendapatkan akurasi, presisi, dan recall sebesar 100%. Hasil sentimen positif diketahui masyarakat akan selalu mendukung dan menginginkan kebebasan untuk Rizieq, untuk sentimen negatif diketahui banyak masyarakat kecewa terhadap Rizieq mengenai kebohongan hasil tes swabnya.

Kata kunci: Asosiasi Teks, Habib Rizieq, Support Vector Machine, Text Mining, Twitter.

Abstract

At the end of 2020, Habib Rizieq's return to Indonesia drew criticism from the public for causing crowds during the Covid-19 pandemic. News and opinions about Habib Rizieq fill internet platforms, including Twitter. The researcher wants to classify the opinion text data of Habib Rizieq's return from Twitter into positive and negative sentiments using the Support Vector Machine method. Opinion data comes from Twitter, so the data is analyzed by text mining through the preprocessing stage. The SVM classification of unbalanced data between positive and negative classes resulted in 95.06% accuracy with a negative class precision value of 84% and better than 72% recall, in the positive class the precision value was 96% less than 2% of recall 98%. While the SVM classification with the oversampling method gets 100% accuracy, precision, and recall. The results of positive sentiments are known that the public will always support and want freedom for Rizieq, for negative sentiments it is known that many people are disappointed with Rizieq regarding the lies of his swab test results.

Keywords: Habib Rizieq, Support Vector Machine, Text Association, Text Mining, Twitter.

INTRODUCTION

At the end of 2020, there was news that managed to capture the attention of the Indonesian people, the news related to the return of the Grand Imam of the Islamic Defenders Front (FPI) to Indonesia on November 10, 2020, after 3.5 years living in Saudi Arabia. It is known that the figure of Grand Imam Habib Rizieq founded an Islamic mass organization known as FPI on August 17, 1998 (Indra, 2021). During his tenure as High Priest of the FPI, Habib has been involved in many cases that have led to him being dragged into legal action. After his return to Indonesia at the end of 2020, Habib Rizieq has assumed the status of a suspect due to 3 different cases, including the crowd at the airport when he arrived in Indonesia, the violation of health protocols in Megamendung, and his action to block the Bogor City Covid-19 Task Force at Ummi Hospital (Saputra, 2018).

The various cases experienced by Habib Rizieq caused his return to Indonesia to trigger the pros and cons of the community. The news regarding Habib Rizieq on the case he experienced after returning to Indonesia became news that received quite a lot of negative comments from the public. His return has become a topic of conversation among Indonesians, both in the real world and on social media (Qurtuby, 2018).

One of the social media that is still accessed by some Indonesians and is an effective medium to accommodate public opinion is Twitter. The Ministry of Communication and Information stated that Indonesia is in the top 5 countries with the largest Twitter users in the world (Riyanto, 2020). The number of users and the ease of access on Twitter in the delivery of opinions will produce a lot of opinion data that can be used as research material, assessment, and evaluation of all cases that happened to Habib Rizieq. The process carried out in generating information from opinion data is by doing sentiment analysis which in its processing will separate opinion data into positive or negative sentiment classes, after which conclusions can be drawn on factors that are often discussed in these opinions.

In this study, data handling is not balanced with oversampling. This was done not only because it wanted to compare with unbalanced data, but it needed to be done to get better and more accurate classification results and to minimize misclassification due to poor understanding of minority data. The oversampling method was chosen because this method is good at duplicating minority data in such a way and making the minority class have a larger sample than before so that it can balance its numbers with the majority class.

So the idea emerged to analyze the opinion of Habib Rizieq's return to Indonesia where opinion data was taken from Twitter social media using scrapping techniques. The data is then classified into positive and negative reviews using the Support Vector Machine (SVM) algorithm. The SVM method is a well-known and powerful tool for the classification of vectors of real-valued features (V. N. Vapnik, 1999) and is good for classifying opinion data obtained from twitter because the SVM method will classify data based on the assessment category they have to find out whether the opinion belongs to the positive or negative sentiment class.

RESEARCH METHODS

Types of research

This study uses qualitative data in the form of text in the form of Habib Rizieq's opinion data on Twitter.

Research Target / Subject

The population used for this research is all public opinion on Twitter regarding the issue of Habib Rizieq's return to Indonesia and the accompanying cases with the keywords #HabibRizieq and #IBhrs. The sample used in this study is all reviews on tweets with #HabibRizieq and #IBhrs from 12-18 January 2021 and 23 February - 1 March 2021, respectively, 616 and 600 tweets of data were obtained.

Data, Instruments, and Data Collection Techniques

This study uses qualitative data in the form of opinion text in tweets. The primary data was obtained by the author through scrapping techniques and the data from Twitter users taken from the Twitter API and the process assisted by software RStudio 3.6.0.

Data analysis technique

In this study, it is known that the data is obtained through the user's Twitter account through the Twitter API. From this data, 1216 data have been obtained based on the keywords #ibhrs and #habibrizieq, after that, the data can be further simplified and then labeled as positive or negative tweets. Then for the SVM method, testing and training data are needed, so for this research case, the example uses a comparison of training and testing data of 80:20, 70:30, and 60:40, in this case, the positive and negative class data are not balanced so the data needs to be processed using The new oversampling is then classified using SVM and compared with the data before the data imbalance is handled using oversampling. The classified data are then evaluated with a confusion matrix to see the performance of the classification results that have been carried out. The flowchart has been explained in Figure 1.

Research Stages

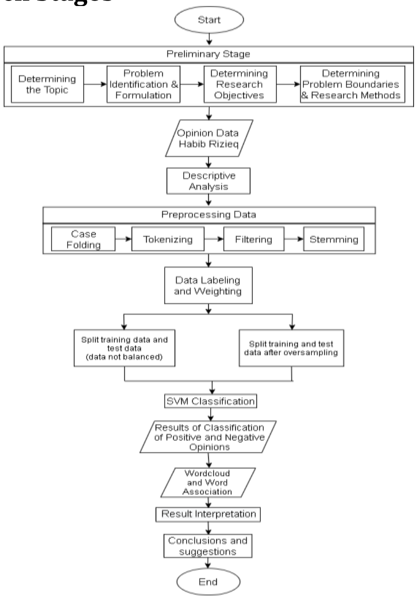
Start
Preliminary Stage
Determining the Topic
Problem Identification & Formulation
Determining Research Objectives
Determining Problem Boundaries & Research Methods
Opinion Data Habib Rizieq
Descriptive Analysis
Preprocessing Data
Case Folding
Tokenizing
Filtering
Stemming
Data Labeling and Weighting
Split training data and test data (data not balanced)
Split training and test data after oversampling
SVM Classification
Results of Classification of Positive and Negative Opinions
Wordcloud and Word Association
Result Interpretation
Conclusions and suggestions
End

Figure 1. Research Flow Chart

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v4i1.262

Preprocessing Stages

Text preprocessing is an early stage of text mining. Referred to as the initial stage by function, this stage of text pre-processing is used to prepare the data and convert it from unstructured data to structured data (Kannan & Gurusamy, 2014). The following are the stages in text preprocessing (Manning, 2009):

1. Case Folding

At the case folding stage, it changes the capital letters in the data to lowercase letters (lowercase).

2. Tokenizing

Tokenizing is a process to parse text, which was originally in the form of a sentence, broken down per word.

3. Filtering

The process of taking the important words needed in the analysis is called filtering. The results of this filtering are in the form of selected words that have been filtered, and those that are connecting words and are not desired by the author will be deleted or removed.

4. Stemming

Stemming is the stage of changing each word from the filtering results that contain affixes into root words.

Support Vector Machine

Support Vector Machines (SVM) is a data algorithm that is growing in popularity in the machine learning community (Yang et al., 2013). SVM is a machine learning method that has the principle of Structural Risk Minimization (SRM), this principle has the aim of separating classes in the algorithm by finding the best hyperplane in the input space (C. C. V. Vapnik, 1995). The best hyperplane is generated by maximizing the minimum margin of the two groups. SVM has a good basic approach in building models so it will have good generalization properties than other methods (V. N. Vapnik, 1999). SVM has the principle of being a linear classifier, but it has begun to be developed so that it can be used in non-linear problems with the kernel trick concept in high-dimensional workspaces (Jangid et al., 2016). The bounding plane hyperplane for class +1 has the following equation in equation (C. C. V. Vapnik, 1995):

$$w.x_i + b \geq +1, untuk\ y_i = +1$$

The equation for hyperplane class -1:

$$w.x_i + b \leq -1\ untuk\ y_i = -1$$

It is known that w is the normal plane and b is the position of the plane relative to the coordinate center. The distance of the boundary plane based on the formula for the distance of the line to the center point is formulated $\frac{1-b-(-1-b)}{w} = \frac{2}{|w|}$. The formula used to search for the best dividing plane with a large margin value in the case of linear classification in primal space is shown in the equation below (Cai et al., 2018):

$$min_w\ ;\ w = \frac{1}{2}|w|^2 \rightarrow min\frac{1}{2}|w|^2$$
$$y_i(w.x_i + b) \geq 1$$

Minimizing objective function $\frac{1}{2}|w|^2$ taking into account the constraint on $y_i(w.x_i + b) \geq 1$.

Oversampling

Imbalanced data is an unbalanced condition that occurs in the classification class. According to (Chawla et al., 2002) the problem of data imbalance can be handled by several methods, including oversampling in the minority class, undersampling in the majority class, and combining the two with a method known as the Synthetic Minority Over Sampling Technique (SMOTE). Oversampling is a technique used to balance data by generating data. The data in this study is not balanced between positive and negative classes, namely the example of comparison data 854:119 or 747:104 where the positive class is many times larger than the negative class, so the oversampling technique is carried out to refine the data to make it more balanced (He et al., 2008).

RESULTS AND DISCUSSION

Descriptive Analysis Of Habib Rizieq Opinion Data

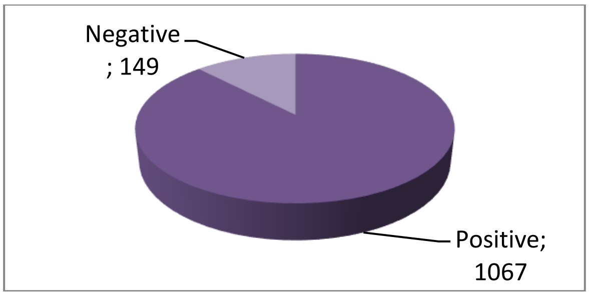

Figure 2. Number of Opinions About Habib Rizieq

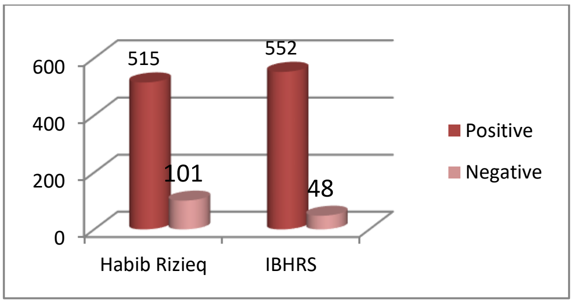

Figure 3. Opinion About Habib Rizieq on #HabibRizieq and #IBHRS

Based on Figure 2, it is known that the number of opinion data about Habib Rizieq used in the study was 1216 tweets with 1067 positive tweets and 149 negative tweets. While in Figure 3 it is explained that in the hashtag Habib Rizieq taken in January, it is known that there are 515 positive opinions and the remaining 101 are negative opinions. In the second hashtag, namely ibhrs, which was obtained in February, 552 positive opinions were obtained and the remaining 48 were negative opinions, there is a very big difference between positive and negative opinions from the second hashtag ibhrs.

Pre-Processing Process

In converting unstructured data into structured data, of course, it is necessary to carry out a preprocessing stage, in this stage consisting of case folding, tokenizing, filtering, and stemming.

Tweet: Pengakuan Habib Rizieq Berbeda dengan Keterangan Aziz Yanuar #HabibRizieq via @jpnncom https://t.co/CPlFdAVYsQ JADI... https://t.co/s0Wl7Yz8FL
Case Folding: pengakuan habib rizieq berbeda dengan keterangan aziz yanuar habibrizieq via jadi
Tokenizing and filtering: pengakuan berbeda keterangan aziz yanuar via
Stemming: aku beda terang aziz yanuar

Sentiment Class Labeling

After doing the preprocessing stage, the next step is modeling the classification process to determine a sentence into members' positive or negative classes based on the calculation value classifier formula. It will be said to be positive if the number of positive words is reduced by the number of negative words the result is more than zero, and otherwise like the calculation example in Table 2. The overall results of the classification data can be seen in Table 1.

Table 1. Opinion Labeling Results

Sentiment	#habibrizieq	#IBhrs	Total
Positive	515	552	1067
Negative	101	48	149

Sentiment Score

The structured data is processed and the score is calculated based on the positive and negative word dictionary used to label the sentiment class. The formula for calculating the sentiment score of an opinion is:

$score = \sum positive\ word - \sum negative\ word$

An example of calculating sentiment scores is shown in Table 2.

Table 2. Example Of Calculating Sentiment Score

Tweet	Positive Word	Negative Word
bohong covid revolusi akhlak bahlul	-	Bohong Bahlul
Total	0	2

Calculation	Skor = 0-2
	Skor = -2	

Training Data And Test Data

Training data is needed in the classification algorithm to form a classifier model. The training data is nothing but used to determine whether the machine is good or bad in understanding data patterns, the larger the training data used, the better the machine in understanding existing data patterns. While the test data is used to measure the accuracy of the classification carried out by the classifier. In this study, 3 data scenarios were used with data comparisons of 80:20 in Table 3, 70:30 in Table 4, and 60:40 in table 5, but in this study, there were problems with unbalanced data, so that in this study, the unbalanced data was processed and handled with the oversampling method.

1. Comparison of training data with test data

Table 3. Data Comparison 80:20

Label	Total	80:20	
		Training Data	Test Data
Positive	1067	854	213
Negative	149	119	30
Total	1216	973	243

Table 4. Data Comparison 70:30

Label	Total	70:30	
		Training Data	Test Data
Positive	1067	747	320
Negative	149	104	45
Total	1216	851	365

Table 5. Data Comparison 60:40

Label	Total	60:40	
		Training Data	Test Data
Positive	1067	640	427
Negative	149	89	60
Total	1216	730	486

2. Handling imbalanced data results in a more balanced data comparison as shown in Table 6.

Table 6. Data sharing after oversampling.

Scenario	Class	Real Data	Oversampling Data
1	P	854	854
	N	119	746
2	P	747	747
	N	104	653
3	P	640	640
	N	89	560

In Table 6 above scenario 1 shows a comparison of training and test data of 80:20, scenario one shows a ratio of 70:30, and scenario 3 is a ratio of 60:40. Likewise, the code for class P indicates a positive class and N indicates a negative class, it can be seen that the data comparison between positive and negative classes before oversampling has so much difference. In contrast to the comparison results after handling unbalanced data with oversampling, the results of the comparison of positive and negative classes are not much different.

SVM Classification

The following are the results of the classification with the Support Vector Machine algorithm with the data scenarios that have been determined.

1. The following Table 7 is the result of the classification with the distribution of training data and test data of 80:20.

Table 7. SVM Data Classification Results in 80:20

Prediction	Actual		Precision
	Negative	Positive	
Negative	21	8	0.84
Positive	4	210	0.96
Recall	0.72	0.98	
Accuracy		0.9506	

Based on the results in table 7, there are 218 positive opinions classified with SVM, it is known that 210 positive opinions are classified in the positive category correctly so that the recall value for positive opinion data is 98% and with a value of information accuracy (precision) of 96%. A total of 25 negative opinion data classified with SVM is known that there are 21 opinions correctly classified into a negative category, so recall value for negative opinion is 72% and a precision value of a negative opinion is 84%. The accuracy value obtained in the 80:20 comparison scenario for classification with the SVM algorithm is 95.06%.

2. Table 8 shows the results of the classification of training data and test data with a proportion of 80:20 using oversampling.

Table 8. 80:20 Data Classification Results with Oversampling

Prediction	Actual		Precision
	Negative	Positive	
Negative	30	0	1.000
Positive	0	213	1.000
Recall	1.000	1.000	
Accuracy		1.000	

Based on the classification using the oversampling method in table 8 above, it is known that from a total of 213 positive opinion data all classified correctly into the positive class, so that they have a recall and precision value of 100%. As well as the negative class, a total of 30 data were all correctly classified into the negative class, so that the negative class also had a recall and precision value of 100%. Due to the perfect classification accuracy, the classification using the oversampling method gets an accuracy value of 100%.

Word Cloud and Word Association

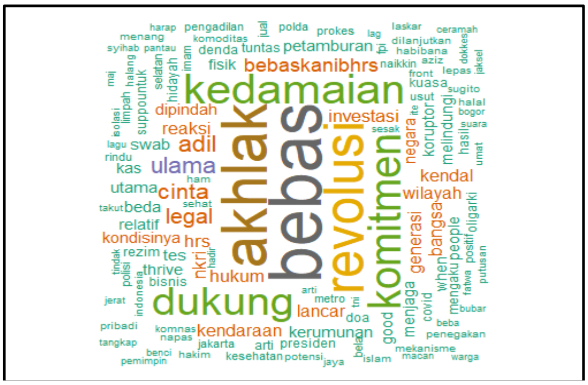

Figure 4. Word Cloud of Positive Class

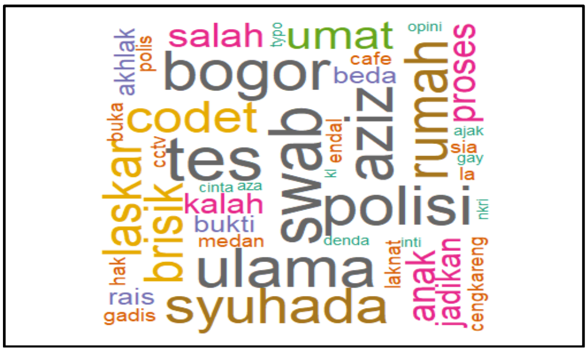

Figure 5. Word Cloud of Negative Class

Based on Figure 4 above, it is clear that the results of the visualization of opinions regarding Habib Rizieq in the positive class. The larger the size of the word in the image, the higher the number of times that word appears in the opinions of Twitter users. Most positive opinions about Habib Rizieq are associated with the word "free", followed by the words "revolution", "morals", "peace", "commitment", "support", "fair", "love", "ulama". Then for the negative words listed in Figure 5, here are some words that often appear or are used in opinions, including "swab", "police", "ulama", "aziz", "bogor", and so on.

The results of word associations in positive classes are often associated with support from the community for Habib Rizieq, hopes from the community for Habib Rizieq to be free from punishment, the moral revolution movement, legal justice for Habib Rizieq, and love for Habib Rizieq. However, in the negative sentiment class that was widely discussed, namely the rejection of the swab test and Habib Rizieq's lies about his swab test results, public disappointment with the police officers who killed FPI soldiers, accusations of giving Habib Rizieq poison to food from the public to the police while in the cell. prisoners and the community agrees with the decision of a minimum of 10 years in prison for Habib Rizieq.

CONCLUSIONS AND SUGGESTIONS

Conclusion

The conclusion that can be drawn is that the best classification is in the data scenario with a ratio of 80:20 without oversampling having an accuracy of 95.06% and with oversampling 100%. The positive word association shows the public's desire for justice and freedom for Habib Rizieq while the negative one shows the public's disappointment with Habib regarding the case of rejecting the swab results and disappointment with the police that killed FPI soldiers.

Suggestion

We recommend that you use another algorithm as a comparison tool from the Support Vector Machine algorithm. Then taking data for a longer period and increasing the number of hashtags will be good for research results to be more representative.

REFERENCES

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, *300*, 70–79. https://doi.org/10.1016/j.neucom.2017.11.077

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*(June), 321–357. https://doi.org/10.1613/jair.953

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *Proceedings of the International Joint Conference on Neural Networks*, *3*, 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

Indra, P. A. (2021). Muhammad Rizieq Shihab. In *Viva.Co.Id*. https://en.wikipedia.org/wiki/Muhammad_Rizieq_Shihab

Jangid, B. M., Jadhav, C. K., Dhokate, S. M., & , Grish M.Jadhav, P. G. M. B. (2016). Proposed Stemming Algorithm for Hindi Information Retrieval. *International Journal of Innovative Research in Computer and Communication Engineering (An ISO Certified Organization)*, *3297*(6), 11449–11455. https://doi.org/10.15680/IJIRCCE.2016

Kannan, S., & Gurusamy, V. (2014). *Preprocessing Techniques for Text Mining (PDF Download Available)*. *October 2014*.

Manning, C. D. (2009). Introduction to Modern Information Retrieval (2nd edition). *Library Review*, *53*(9), 462–463. https://doi.org/10.1108/00242530410565256

Qurtuby, S. al. (2018). *sumanto*. https://www.dw.com/id/mayoritas-publik-menolak-rizieq-shihab-kembali-ke-indonesia/a-46573605

Riyanto, A. D. (2020). Indonesia Digital report 2020. *Global Digital Insights*, 43. https://datareportal.com/reports/digital-2020-indonesia?rq=digital-2020-indonesia

Saputra, A. R. I. (2018). *Ari saputra* (p. 2595). https://news.detik.com/berita/d-5328884/3-status-tersangka-habib-rizieq-di-3-kasus

V. Vapnik, C. C. (1995). Photonic neural networks and learning machines the role of electron-trapping materials. *IEEE Expert-Intelligent Systems and Their Applications*, *7*(5), 63–72. https://doi.org/10.1109/64.163674

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, *10*(5), 988–999. https://doi.org/10.1109/72.788640

Yang, Z. M., He, J. Y., & Shao, Y. H. (2013). Feature selection based on linear twin support vector machines. *Procedia Computer Science*, *17*, 1039–1046. https://doi.org/10.1016/j.procs.2013.05.132