Journal of
Information Systems Engineering
and Business Intelligence

Vol.7, No.2, October 2021
Available online at: http://e-journal.unair.ac.id/index.php/JISEBI

Sentiment Analysis Towards Kartu Prakerja Using Text Mining with Support Vector Machine and Radial Basis Function Kernel

Belindha Ayu Ardhani[1)], Nur Chamidah[2)*], Toha Saifudin[3)]

[1,2,3)]*Department of Mathematics, Faculty of Science and Technology, Universitas Airlangga, Indonesia*
Dr. Ir. H. Soekarno St., Mulyorejo, Surabaya
[1)]belindha.ayu.ardhani-2017@fst.unair.ac.id, [2)]nur-c@fst.unair.ac.id, [3)]tohasaifudin@fst.unair.ac.id

Abstract

Background: The introduction of Kartu Prakerja (Pre-employment Card) Programme, henceforth KPP, which was claimed to have launched in order to improve the quality of workforce, spurred controversy among members of the public. The discussion covered the amount of budget, the training materials and the operations brought out various reactions. Opinions could be largely divided into groups: the positive and the negative sentiments.
Objective: This research aims to propose an automated sentiment analysis that focuses on KPP. The findings are expected to be useful in evaluating the services and facilities provided.
Methods: In the sentiment analysis, Support Vector Machine (SVM) in text mining was used with Radial Basis Function (RBF) kernel. The data consisted of 500 tweets from July to October 2020, which were divided into two sets: 80% data for training and 20% data for testing with five-fold cross validation.
Results: The results of descriptive analysis show that from the total 500 tweets, 60% were negative sentiments and 40% were positive sentiments. The classification in the testing data show that the average accuracy, sensitivity, specificity, negative sentiment prediction and positive sentiment prediction values were 85.20%; 91.68%; 75.75%; 85.03%; and 86.04%, respectively.
Conclusion: The classification results show that SVM with RBF kernel performs well in the opinion classification. This method can be used to understand similar sentiment analysis in the future. In KPP case, the findings can inform the stakeholders to improve the programmes in the future.

Keywords: Kartu Prakerja, Sentiment Analysis, Support Vector Machine, Text Mining, Radial Basis Function

Article history: Received 2 May 2021, first decision: 15 June 2021, accepted 23 September 2021, available online 28 October 2021

I. INTRODUCTION

Kartu Prakerja (Pre-employment Card) Programme (KPP), one of flagship programmes the current government of Indonesia rolled out in 2020, became a controversy in various media soon after it was launched. The programme was designed before the COVID-19 pandemic, aimed at improving the quality of workforce through upskilling and training [1]. The aspects scrutinised include the amount of budget, the training materials and the operations. Criticisms targeted the technical implementation, which was deemed off-track. Therefore, in response to the COVID-19 pandemic, the programme was adjusted to become a social safety net, referring to the fact that over 1.2 million workers from 74,439 companies in both the formal and informal sectors were laid off [2]. The government estimated that up to 3.78 million Indonesians would fall into poverty and 5.2 million would lose their jobs due to the pandemic [3]. KPP was expected to play a strategic role in the pandemic era, so it is necessary to improve the quality of services and facilities provided by evaluating the public opinions. These opinions can be largely classified into two groups: the positive and negative sentiments. Positive sentiments are identified from the perceived satisfaction and the support given. These include sentiments about the various types of training offered; the incentives provided to mitigate the impact of the pandemic; the high number of participants accepted in each batch of training; the usefulness of the training provided in improving skills; the suitability of the curriculum to the actual needs in the industry; the counselling and consultation. Meanwhile, the negative sentiments are identified through the complaints, which include: lack of quality and high cost of training

* Corresponding author

ISSN 2443-2555 (online) 2598-6333 (print)
doi: http://dx.doi.org/10.20473/jisebi.7.2.119-128

materials; late disbursement of incentives; ineligible participants being admitted; hoaxes related to the pre-employment card link; unstable server; difficulties in uploading registration files; and difficulty to report problems experienced to the contact persons.

The automated method used in the sentiment analysis is through text mining from the opinions circulated on social media. Twitter has a number of daily active users that increased by 34% to 186 million in the second quarter of 2020 [4]. It is also a text based social media platform. Text mining is an appropriate method for identifying and extracting textual information [5], including from Twitter. The implementation is done using Support Vector Machine (SVM), which is part of supervised learning to perform non-probabilistic binary classification. SVM classifies text into two fields with principle of maximising the margin or distance between the bounding planes [6]. The algorithm is chosen in this research based on the advantages of it. Darnag, et al [7] explained that the advantage of SVM is that it adopts Structure Risk Minimization (SRM) principle, which gives smaller generalisation errors than the Empirical Risk Minimization (ERM) results in conventional neural networks. That is why many researchers have reported that SVM is very accurate to use for text classification [8]. In the implementation, nonlinear classification of SVM is more widely applied to various cases. This is due to the fact that generally data in real life are rarely linear. For this reason, SVM is modified by including kernel functions [9]. Between four types of kernel in SVM—linear kernel, polynomial kernel, Radial Basis Function (RBF) kernel, and sigmoid kernel—it is known that RBF is the best kernel in most cases.

In summary, this research contributes to the automation of sentiment classification about KPP as recommendation to simplify the evaluation process by Ministry of Manpower and related institutions. The automation using text mining in this research could expedite the improvement needed in order to make the programme more effective. The use of appropriate analytical method can produce an accurate classification and the robust findings can subsequently inform the relevant stakeholders in the decision-making process.

II. Literature Review

Previous studies have implemented text mining for sentiment analysis on Twitter. One of them is research from P. H. Prastyo, et al [10]. The tweets used were from 23 March 2020 to 14 May 2020. The total data were 2,203 tweets for the general aspects, which were scraped using keywords "#COVID-19indonesia; COVID-19 di Indonesia; and penanganan pemerintah terhadap COVID19", as well as 1,941 tweets on the economic aspects, which were scraped using keywords "Dampak ekonomi karena COVID-19 di Indonesia, penanganan ekonomi COVID-19 di Indonesia". Regarding this, the method used is svm with normalised poly kernel to classify tweets responding to the Indonesian government's handling of COVID-19. Their analysis achieved the highest performance in average accuracy, precision, recall and F-measure respectively with the value of 82.00%, 82.24%, 82.01% and 81.84% [10].

Meanwhile, P. Dellia and A. Tjahyanto [11] conducted tax complaints classification on Twitter using Naive Bayes, SVM and decision tree. The total tweets used were 1,001 tweets with 758 tweets labelled as no-complaints and 243 tweets labelled as complaints. Those were scraped using keyword "@DitjenPajakRI" and "@kring_pajak". The results showed that SVM had the highest value of F-measure compared to Naive Bayes and decision tree, which were respectively 89.3%, 85.6% and 76.9%.

Shofiya and Abidi [12] also conducted a sentiment analysis using Twitter data. The topic is about COVID-19 Social Distancing in Canada. They used 629 tweets, which were classified into positive, negative and neutral sentiments. Using SVM, they showed that the performance evaluation has an accuracy of 71%. It increased to 81% when only positive and negative sentiment polarity were used. It was also observed that reducing test data by 10% increased the accuracy to 87%. Another study [13] classifying 833 tweets with the keyword "Corona Virus" into positive, negative and neutral sentiments by using Naive Bayes, SVM, and K-Nearest Neighbour (KNN) methods showed that the best performance result was produced by the SVM with an accuracy value of 76.21%, a precision value of 78.04%, and a recall value of 71.42% [13].

III. Methods

The data used in this research are tweets on KPP. They were classified using SVM RBF kernel with R software. The research variables used consist of response variable and predictor variable. The former is class of sentiment with two categories, positive sentiment and negative sentiment; and the latter is public opinions regarding KPP. The steps of analysis are presented in Fig. 1.

A. Data Crawling

Tweets on KPP were retrieved by using Application Programming Interface (API). The data were taken using keyword "prakerja" with total of 500 tweets in the period of July to October 2020.

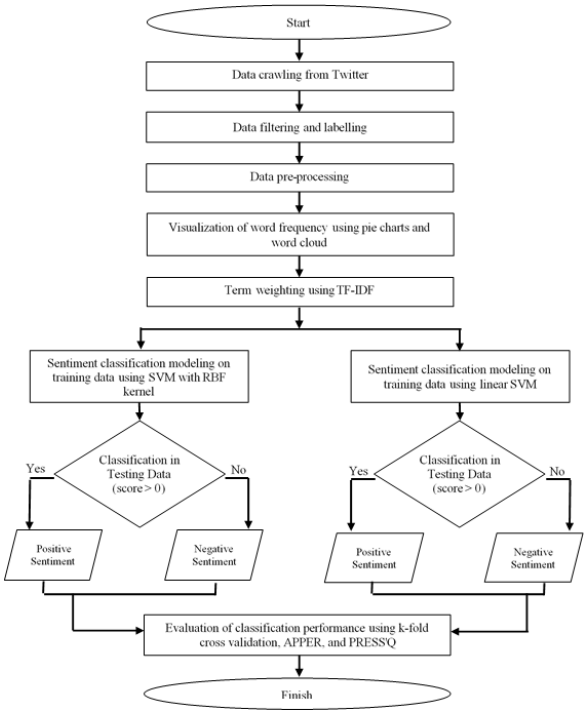

Fig. 1 Analysis Steps Flowchart

B. Data Filtering and Labeling

Data filtering process aims to remove tweets that contain pornography, threats or spam. This is followed by data labelling, which in this research was carried out by one person. The label consists of positive and negative classes. The results of the labelling process in this research are shown in Table 1.

TABLE 1
LABELING DATA RESULTS

Tweet	Sentiment
Keren nih yg dapet insentif dari Prakerja pada dipake buat usaha. <U+0001F44F><U+0001F3FD> tepat guna dananya.	Positive
Sulit sekali ya daftar PraKerja.	Negative
Di janjian terus menerus, hal Insentifnya selalu telat,,, nasibmu #Prakerja @Bim_Bim_Slank @iwanfals	Negative
@CNNIndonesia Program dana prakerja gak jelas .. Gak tepat sasaran .. Banyak dimanfaatin org yg tau cara ngambil y .. Dan mreka semua pekerja bahkan ekonomi mnengah ke atas yg mnikmati y .. Heran https://t.co/4qepDQ2jJi	Negative
bingung milih pelatihan yg mana d program kartu prakerja, pelatihannya bagus2.. bisa d jadiin bahan buat jadi teacherpreneur, Aamiin~	Positive

TABLE 2
PRE-PROCESSING DATA RESULTS

Stage	Definition of Stage	Result of Stage
Cleansing	Cleansing the document from unnecessary words or characters, such as html, mention, hashtag (#), URL, numbers, punctuation, and emotion icon.	Program dana prakerja gak jelas Gak tepat sasaran Banyak dimanfaatin org yg tau cara ngambil y Dan mreka semua pekerja bahkan ekonomi mnengah ke atas yg mnikmati y Heran
Case Folding	Uniformity of all Latin letters A to Z, which are capitalised to lowercase letters "a" to "z", while eliminating excess letters into one letter.	program dana prakerja gak jelas gak tepat sasaran banyak dimanfaatin org yg tau cara ngambil y dan mreka semua pekerja bahkan ekonomi mnengah ke atas yg mnikmati y heran
Spelling Normalization	Correcting words that do not meet the rules, both in the form of spelling and abbreviations by using slang word dictionary. The slang word dictionary was made manually, which is specific for KPP.	program dana prakerja gak jelas gak tepat sasaran banyak dimanfaatkan orang yang tahu cara ambil ya dan mereka semua pekerja bahkan ekonomi tengah ke atas yang menikmati ya heran
Stopword Removal	Removal of irrelevant and meaningless words in the document by using a stop-word dictionary.	program dana prakerja jelas tepat sasaran banyak dimanfaatkan tahu cara ambil pekerja ekonomi menikmati heran
Tokenizing	Splitting documents into pieces of words to become entities that have value in the text document matrix. In this step, space is the delimiter for tokenisation.	program dana prakerja jelas tepat sasaran banyak dimanfaatkan tahu cara ambil pekerja ekonomi menikmati heran
Stemming	Extraction to get basic words to reduce word variations. This research used Indonesian-specific stemmer.	program dana prakerja jelas tepat sasar banyak manfaat tahu cara ambil kerja ekonomi nikmat heran

C. Pre-processing Data

Pre-processing data aims to prepare data for pattern extraction, considering that the data were unstructured. It consisted of cleansing, case folding, spelling normalisation, stop-word removal, tokenising and stemming [14]. Table 2 shows the result for one of the tweets obtained in Table 2.

D. Visualisation of Word Frequency

Visualisation by using descriptive statistics aims to understand the overview of the public opinions. Besides using pie charts, word cloud—a simple tool to identify the focus of written material [15]—was also used.

E. Term Weighting Using TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) is word weighting, which aims to make the data eligible for processing and calculation to create a model [16]. TF in detail is the frequency of words in each document [17]. Meanwhile, IDF shows how important a word is in the entire document. This representation gives more weight to words that occur less frequently than the words that occur more frequently. For this reason, TF-IDF is useful for balancing the weight between the words that most often used and words that are rarely used [18]. IDF is defined in (1).

$$idf_{ij} = log\left(\frac{D}{df_i}\right) \tag{1}$$

where:
idf_{ij} : IDF value of the i^{th} word in the j^{th} document
D : Total number of documents in a collection (corpus)
df_i : Number of documents containing the i^{th} word in the corpus

Thus, mathematically TF-IDF can be formulated by (2) as follows.

$$w_{ij} = tf_{ij} \times idf_{ij} \tag{2}$$

where w_{ij} is the i^{th} word weight in the j^{th} document; and tf_{ij} is the frequency of the i^{th} word in the j^{th} document.

F. Classification Using SVM RBF

After conducting pre-processing data and term weighting, the data were divided into 80% data for training and 20% data for testing with five-fold cross validation. The model formed by training data is used to classify the testing data. It is analysed using SVM.

SVM is a method used to perform non-probabilistic binary classification. Conceptually, it uses the results of training model to find the best hyperplane in the classifying data [18]. Hyperplane in SVM is defined as a separating function between two classes. SVM works to find optimal solutions for these functions. The best hyperplane between the two classes is determined based on the maximum margin measurement. Margin is the distance between the data points that enter the positive class and the data points that enter the closest negative class around the hyperplane. This data point is also known as a support vector, which is the outermost data object closest to the hyperplane. In this case, support vectors are used as the best hyperplane [19]. In general, the decision function that determines the classification class is shown in (3).

$$f(\boldsymbol{x}_k) = \sum_{i=1}^{ns} \alpha_i y_i \boldsymbol{x}_i \cdot \boldsymbol{x}_k + b \tag{3}$$

where *ns* is the number of support vectors; α_i is the lagrange coefficient; $\boldsymbol{x}_i$ is a support vector; $\boldsymbol{x}_k$ is the new input data (document); *b* is bias.

Equation (3) is used in the hard margin SVM, where data can be separated linearly. However, data in reality are filled with noise (outlier) and they are difficult to be separated linearly. In this case, the hard margin SVM classification cannot provide optimal results if there are outliers in the data. One solution to this nonlinear data classification is kernel trick. Thus, the decision function can be presented in (4) [20].

$$f(\boldsymbol{x}_k) = \sum_{i=1}^{ns} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}_k) + b \tag{4}$$

SVM has several types of kernels—linear kernel, polynomial kernel, Gaussian kernel or Radial Basis Function (RBF), and sigmoid kernel. RBF is recommended one to use, referring to its ability to produce a hyperplane that can separate classes by adjusting nonlinear data patterns. RBF becomes a function whose value only depends on the distance from several points [21]. The RBF kernel is defined in (5) below.

$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\gamma \|\boldsymbol{x} - \boldsymbol{x}'\|^2\right) \quad (5)$$

There is a gamma parameter that plays a role for measuring similarity between data points. The small gamma value indicates that the support vector has a big effect on vector class determination of the input data even though it is far away. For this reason, if the gamma value is too small, the model behaves like a linear SVM. Meanwhile, a large gamma value indicates that the support vector has no effect on vector class determination of the input data broadly. That is why if the gamma is too large, the model will be affected by each support vector.

G. Evaluation of Classification Performance

The evaluation of algorithm performance in the classification results is carried out using several methods. One of them is k-fold cross validation. It is used to test the validation of classifier by partitioning the data into training data and testing data. The use of this method can reduce bias in sampling because the data is divided randomly into several (k) parts for training using several parts; and tested in other sections. In relation to this process, the final accuracy is the average accuracy of the number of processes [22]. The illustration of data partition using k-fold cross validation is showed in Fig. 2.

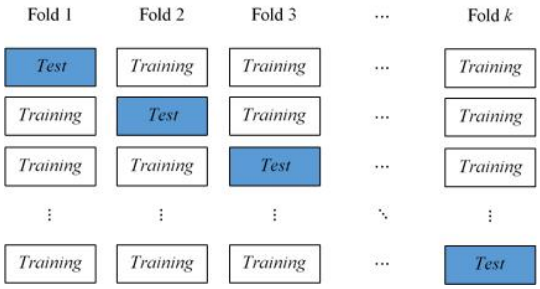

Fig. 2 K-Fold Cross Validation

Other method to evaluate the performance is confusion matrix. There are many indicators measured in the confusion matrix, which can be seen as in Table 3.

TABLE 3
CONFUSION MATRIX FOR EVALUATION

		Actual Value	
		Positive	**Negative**
Predictive Value	**Positive**	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Apparent Error Rate (APPER), which shows probability of error in classifying objects, is calculated using (6).

$$APPER = \frac{FP+FN}{TP+FP+FN+TN} \quad (6)$$

while the accuracy is calculated using (7) as follows.

$$Accuracy = 1 - APPER \quad (7)$$

In addition, sensitivity, specificity, positive predictive value and negative predictive value are calculated respectively using (8), (9), (10), and (11) [23].

$$Sensitivity = \frac{TP}{(TP+FN)} \quad (8)$$

$$Specificity = \frac{TN}{(TN+FP)} \quad (9)$$

$$Positive\ Predictive\ Value = \frac{TP}{(TP+FP)} \quad (10)$$

$$Negative\ Predictive\ Value = \frac{TN}{(TN+FN)} \quad (11)$$

IV. RESULTS

A. Descriptive Analysis of the Public Opinions on KPP

Based on the data obtained, there were 500 tweets from thousands of tweets, which contain opinions. The public opinions actually show mixed sentiments. Fig. 3 below shows that 300 were negative sentiments; while the remaining ones were positive sentiments. These negative sentiments indicate that there are several aspects of KPP that need to be improved.

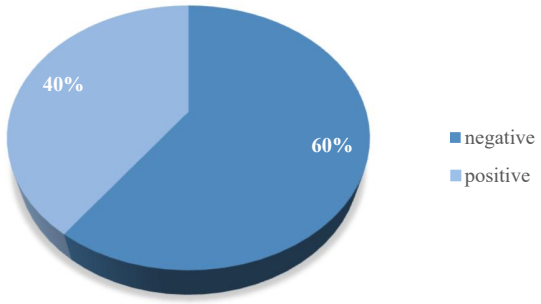

Fig. 3 Percentage of Public Opinions Based on Sentiment Classification

Then, if identified, the opinions on sentiment classifications have a number of meaningful words that appear frequently. Those were based on feature selection using TF-IDF and then they were visualised using word cloud. The more often a word appears in the analysed text, the greater the word will appear in the result of word cloud [15]. The visualisation of these is presented in Fig. 4.

Fig. 4 Word Cloud of Public Opinions Containing Positive Sentiment

Fig. 5 Word Cloud of Public Opinions Containing Negative Sentiment

The words showed in Fig. 4 and Fig. 5 are the result of the pre-processing, which includes removing meaningless words (stop-word removal), such as conjunctions. Fig. 4 was created based on the opinions with a positive label, while Fig. 5 was created based on the opinions with a negative label. For this reason, there are several words from Fig. 4 that also appear in Fig. 5 and vice versa. It happens because those words are important terms that help clarify the context of an opinion. Therefore, the opinion can be included in the positive or negative class.

B. Public Opinion Classification Using Support Vector Machine with Radial Basis Function Kernel

The classification of training data is carried out to create sentiment classification model from $\boldsymbol{x}_i$ documents, which consist of word weight resulted by TF-IDF. It aims to train the machine to automatically classify new opinions into sentiment classes. The results of weight using TF-IDF are used as input data for the classification. In this modelling process, the optimal parameters of RBF kernel consisting of cost *C* and gamma γ are determined firstly. The selected optimal parameters and the specified training data are used to produce the lagrange coefficient α_i. Then, α_i value is used to obtain support vectors and parameter *b*, so the model can be formed as a function of classification decisions.

In this case, the optimal RBF kernel parameters are determined using grid search method. This method is done by iterating the training data using different RBF kernel parameters. Exponential addition is used in each iteration. After getting a certain range, it needs to perform parameters in a smaller value range so the optimal parameter is found. In this research, determining the optimal parameters for each fold was carried out with the help of function tune in library package e1071. The optimal parameters chosen have the highest accuracy in classifying the training data in each fold. Table 4 shows the classification performance using the optimal parameters of RBF kernel.

TABLE 4
SVM WITH RBF KERNEL PERFORMANCE IN TRAINING DATA

Fold	Optimal Parameter		Acuracy	Sensitivity	Specificity	Negative Sentiment Predictive Value	Positive Sentiment Predictive Value
	C	γ					
1	4	0.1000	1.0000	1.0000	1.0000	1.0000	1.0000
2	4	0.1500	1.0000	1.0000	1.0000	1.0000	1.0000
3	4	0.0625	0.9975	1.0000	0.9936	0.9959	1.0000
4	8	0.0625	0.9975	1.0000	0.9939	0.9958	1.0000
5	4	0.2500	1.0000	1.0000	1.0000	1.0000	1.0000
Average			**0.9990**	**1.0000**	**0.9975**	**0.9983**	**1.0000**

The modelling of each training data is carried out alternately, so there are five iterations with k-fold cross validation. Based on Table 4, the performance measurement for each fold produces average validation, sensitivity, specificity, negative sentiment predictive value and positive sentiment predictive value. In their respective order, they are 99.90%; 100%; 99.75%; 99.83%; and 100%. The optimal parameters in each fold are used to create classification model. In this regard, the training data for each fold has different model from one another. The calculation in (12) below presents one of the hyperplanes formed for opinion classification in the 5th fold.

$$f(\boldsymbol{x}_d) = \sum_{i=1}^{375} \alpha_i y_i \, K(\boldsymbol{x}_i, \boldsymbol{x}_d) + b$$

$$= \sum_{i=1}^{375} \alpha_i y_i \, exp(-0.25\|\boldsymbol{x}_i - \boldsymbol{x}_d\|^2) + 0.17397 \quad (12)$$

The classification models obtained from the training data were used to classify sentiments on the testing data. Classification prediction in the testing data is done using the predict function. This prediction is also carried out iteratively with 5-fold cross validation according to the training data for each fold. The results of the classification are presented in Table 5.

TABLE 5
SVM WITH RBF KERNEL PERFORMANCE IN TESTING DATA

Fold	Optimal Parameter		Accuracy	Sensitivity	Specificity	Negative Sentiment Predictive Value	Positive Sentiment Predictive Value
	C	γ					
1	4	0.1000	0.8300	0.9500	0.6500	0.8028	0.8966
2	4	0.1500	0.8200	0.8667	0.7500	0.8387	0.7895
3	4	0.0625	0.8200	0.9123	0.6977	0.8000	0.8571
4	8	0.0625	0.8900	0.9048	0.8649	0.9194	0.8421
5	4	0.2500	0.9000	0.9500	0.8250	0.8906	0.9167
Average			**0.8520**	**0.9168**	**0.7575**	**0.8503**	**0.8604**

The results in Table 5 were obtained by entering the testing data into the classification model before corresponding them to the training data. Based on Table 5, an average accuracy value of 85.20% means that the results of the classification prediction in this testing data are close to the actual classification. While the average sensitivity value was 91.68%. This value indicates that the ability of the SVM with RBF kernel is excellent in classifying negative sentiment opinions from all samples that are negative sentiments. On the total actual value of sensitivity, the average specificity value obtained was 75.75%, indicating that the ability of this algorithm is good in classifying positive

sentiment opinions that are actually positive sentiments. Meanwhile, the negative sentiment predictive value, which means that the percentage of negative sentiment prediction that is actually negative showed an average of 85.03%. Regarding this, the average positive sentiment predictive value was 86.04%. On the whole, this indicated a good classification ability in the testing data.

The results of classification on testing data are evaluated to determine the performance of SVM with RBF kernel in the classification of public opinions about KPP. The performance evaluation was carried out using several methods. APPER was the first evaluation method performed on the classification results. APPER was used to determine the proportion of misclassification. Table 6 shows the calculation of the APPER value by taking one of the folds in Table 4. In this case, the 5th fold was chosen for the APPER classification calculation because it has higher performance value than the other folds.

TABLE 6
CONFUSION MATRIX RESULTS OF THE 5TH FOLD TESTING DATA CLASSIFICATION USING SVM WITH RBF KERNEL

		Actual Value		Total
		Positive	Negative	
Predictive Value	Positive	33	3	36
	Negative	7	57	64
Total		40	60	100

Based on Table 6, the sentiments that were classified correctly for the positive class were 33 opinions; and for the negative class were 57 opinions. Meanwhile, the number of opinions that were not classified correctly were 10 opinions. The calculation of misclassification proportion with APPER is presented in (13) below.

$$APPER = \frac{FP+FN}{TP+FP+TN+FN} \times 100\%$$
$$= \frac{10}{100} \times 100\% = 10\% \quad (13)$$

The APPER calculation result in (13) shows that the classification error rate was 10%. Thus, the value of classification accuracy obtained was 90% in accordance with the R output in Table 3.

C. Evaluation of Support Vector Machine Performance with Radial Basis Function Kernel on Sentiment Analysis Results of KPP

Related to the classification results, SVM RBF kernel as nonlinear SVM needs to be compared with linear SVM in this case. This aims to determine which SVM has the best performance to classify opinions on KPP. The comparison between the SVM RBF kernel and linear SVM is done based on the performance indicators from the average of five folds in the testing data used as in Table 7.

TABLE 7
PERFORMANCE COMPARISON OF LINEAR SVM AND SVM WITH RBF KERNEL

Algorithm	Acuracy	APPER	Sensitivity	Specificity	Negative Sentiment Predictive Value	Positive Sentiment Predictive Value
Linear SVM	0.8300	0.1700	0.8799	0.7565	0.8446	0.8106
SVM RBF kernel	0.8520	0.1480	0.9168	0.7575	0.8503	0.8604

Table 7 shows that nonlinear SVM, which is SVM with RBF kernel produces better performance than linear SVM. This is evidenced by the greater value of accuracy, sensitivity, specificity, negative sentiment predictive value and positive sentiment predictive value with smaller APPER. Referring to the pattern of public opinions data which are irregular and difficult to separate using linear line, the use of SVM kernel RBF was the most suitable.

V. DISCUSSION

Social media has now become an integral part of modern life and people tend to voice their opinions about an issue of their concerns. Textual data retrieved from Twitter gives many opportunities to analyse topics by using big volume and velocity. This research focuses on the sentiment classification of KPP. Public opinions can be classified automatically to analyse the success rates of certain governmental programmes. Classification using text mining can alleviate the workload to do such analysis significantly and produce more robust results at the same time.

In the implementation of text mining using SVM, the use of nonlinear SVM is suitable to classify data, where classes cannot be separated linearly in two-dimensional space. The public opinions on KPP data is this kind of data.

The results of analysis using RBF kernel reached high accuracy, which means that this algorithm is effective in classifying opinions into positive or negative sentiment. Furthermore, the comparison results, which aim to analyse performance between the nonlinear SVM and linear SVM, also supported the ability of SVM with RBF kernel. It shows that SVM with RBF kernel yielded better results. It is not only shown through the accuracy value, but also other performance measurements such as sensitivity, specificity, negative sentiment prediction value and positive sentiment prediction value. In relation to this, the SVM with RBF kernel has higher values. It is also supported with smaller APPER value. This is in line with the previous studies [24], which have proven SVM RBF performance. Thus, it can be concluded that the SVM with RBF kernel is successful in classifying public opinions on KPP.

Nonetheless, this study has several limitations. The first limitation is, only one annotator in the labelling process. Because it can create bias, we double checked the labelling results. This check was conducted by the same person. We suggest that there should be more annotators and then the labelling results can be tested with inter-reliability test (Cohen's Kappa). Additionally, there are still words that do not specifically represent a negative or positive sentiment. Thus, the next research should use additional feature selection, such as Chi Square Test so they really reflect either positive or negative opinions. In addition, it is better to use more data and compare he types of kernel function to determine the most suitable kernel for the cases discussed.

VI. Conclusions

Based on the results and discussion, the use of SVM with RBF Kernel results in good performance for sentiment classification about KPP. Thus, it can help to build the automation process during KPP implementation. It is useful to make the analysis of governmental programme easier. Then, the analysis can be a reference to optimise the role and function of the programmes to support the manpower sector and the economy in general.

Author Contributions: *Belindha Ayu Ardhani*: Methodology, Visualization, Data Curation, Software, Writing-Original Draft, *Nur Chamidah*: Conceptualization, Supervision, Investigation, Writing-Review & Editing, Funding acquisition, *Toha Saifudin*: Formal Analysis.

Funding: This research received no specific grant from any funding agency.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] M. A. Iswara, "Indonesia Advances Preemployment Card Launch to Friday to Anticipate Virus Impacts", Retrieved from https://www.thejakartapost.com/news/2020/03/18/indonesia-advances-preemployment-card-launch-to-friday-to-anticipate-virus-impacts.html, 2020.

[2] M. A. Iswara, "1.2 Million Indonesian Workers Furloughed, Laid off as COVID-19 Crushes Economy", Retrieved from https://www.thejakartapost.com/news/2020/04/09/worker-welfare-at-stake-as-covid-19-wipes-out-incomes.html, 2020.

[3] A. W. Akhlas, "Millions to Lose Jobs, Fall into Poverty as Indonesia Braces for Recession", Retrieved from https://www.thejakartapost.com/news/2020/04/14/millions-to-lose-jobs-fall-into-poverty-as-indonesia-braces-for-recession.html , 2020.

[4] J. Goldsmith, "Twitter Active Daily Users Surge 34% To Record 186M In Q2, Revenue Dips, CEO Jack Dorsey Apologizes For Breach", Retrieved from https://deadline.com/2020/07/twitter-active-daily-users-surge-34-to-186m-q2-revenue-dips-19-1202992835/, 2020.

[5] N. Zanini, and V. Dhawan, "Text Mining: An Introduction to Theory and some Applications", Research Matters: A Cambridge Assessment Publication*,* Issue 19, Pages 38–44, 2015.

[6] M. Sugiyama, "Introduction to Statistical Machine Learning", Elsevier:Waltham, 2015.

[7] R. Darnag, B. Minaoui, and M. Fakir, "QSAR Models for Prediction Study of HIV Protease Inhibitors Using Support Vector Machines, Neural Networks and Multiple Linear Regression", Arabian Journal of Chemistry, Vol. 10, Supplement 1, Pages S600–S608, 2017.

[8] R. Moraes, J. F. Valiati, and W. P. G. Neto, "Document-level sentiment classification: An empirical comparison between SVM and ANN", Expert Systems with Applications, Vol. 40, Issue 2, Pages 621–633, 2013.

[9] S. Liu, J. McGree, Z. Ge, and Y. Xie, "Computational and Statistical Methods for Analysing Big Data with Applications", Academic Press:UK, 2016.

[10] P. H. Prastyo, A. S. Sumi, A. W. Dian, A. E. Permanasari, "Tweets Responding to the Indonesian Government's Handling of COVID-19: Sentiment Analysis Using SVM with Normalized Poly Kernel", Journal of Information Systems Engineering and Business Intelligence, Vol. 6, No. 2, Pages 112–122, 2020.

[11] P. Dellia and A. Tjahyanto, "Tax Complaints Classification on Twitter Using Text Mining", IPTEK, Journal of Science, Vol. 2, No. 1, Pages 11–15, 2017.

[12] C. Shofiya, and S. Abidi, "Sentiment Analysis on COVID-19-Related Social Distancing in Canada Using Twitter Data", International Journal of Environmental Research and Public Health, Vol. 18, Issue 5993, Pages 1–10, 2001.

[13] R. Risnantoyo, A. Nugroho, and K. Mandara, "Sentiment Analysis on Corona Virus Pandemic Using Machine Learning Algorithm", Journal of Informatics and Telecommunication Engineering, Vol. 4, No. 1, Pages 86–96, 2020.

[14] G. Miner, J. Elder, A. Fast, T. Hill, R. Nisbet, and D. Delen, "Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications", Academic Press:USA , 2012.

[15] R. Atenstaedt, and C. Singh, "Word Cloud Analysis of BJGP", British Journal of General Practice, Vol. 62, Issue 596, Pages 148, 2012.

[16] Sahana, "Advances in Computational Intelligence". Springer:Switzerland, 2020.

[17] R. Chopra, A. M. Godbole, N. Sadvilkar, M. B. Shah, S. Ghosh, and D. Gunning, "The Natural Language Processing Workshop", Packt: Birmingham, 2020.

[18] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of Sentimental Reviews Using Machine Laerning Techniques", Procedia Computer Science, Vol. 57, Pages 821–829, 2015.

[19] M. Awad, and R. Khanna, "Efficient Learning Machines : Theories, Concepts, Applications for Engineers and System Designers", Apress:New York, 2015.

[20] C. Cortes, and V. Vapnik, "Support-Vector Networks", Machine Learning, Vol. 20, Pages 273–297, 1995.

[21] A. Kowalczyk, "Support Vector Machine Succinctly", Syncfusion:USA, 2017.

[22] M. Asrol, P. Papilo, and F. E. Gunawan, "Support Vector Machine with K-fold Validation to Improve the Industry's Sustainability Performance Classification", Procedia Computer Science, Vol. 179, Pages 854–862, 2021.

[23] A. C. Rencher, "Methods of Multivariate Analysis Second Edition", John Wiley & Sons:USA, 2003.

[24] H. Hong, B. Pradhan, D. T. Bui, C. Xu, A. M. Youssef, and W. Chen, "Comparison of Four Kernel Functions Used in Support Vector Machines for Landslide Susceptibility Mapping: a Case Study at Suichuan Area (China)", Geomatics, Natural Hazards snd Risk, Vol. 8, No. 2, Pages 544–569, 2017.