Proceeding International Conference on Religion, Science and Education (2023) **2**, 761-765

A Comparison of SVM Kernel Functions for Sentiment Analysis of UU TPKS

Ni Kadek Mirah Budayani[1], Isnandar Slamet[2], Sri Sulistijowati Handajani[3]
[1,2,3]Statistics Department, Faculty of Mathematics and Natural Sciences, Sebelas Maret University,
Jl. Ir. Sutami No. 36-A, Kentingan, Surakarta 57126, Indonesia. Tel. +62-271-646994, Fax. +62-271-646655.

Corresponding author
[1]mirahbudayani@student.uns.ac.id

Abstract: The law on the elimination of sexual violence or Undang-Undang Tindak Pidana Kekerasan Seksual (UU TPKS) previously known as draft law on criminal sexual violence got approved in April 12th, 2022 in the plenary meeting of the Indonesia's Parliament. Before the law has been approved, the bills were initiated by the National Commission on Violence Against Women (Komnas Perempuan) in 2012. The passed bills raised many reactions from Indonesian Citizen on social media, specifically Twitter. In this study, we do sentiment analysis on the passes of UU TPKS using 5486 data by keywords UU TPKS and/or #UUTPKS on Twitter. Different kernel functions with different combination of C, gamma, and degree on support vector machine used to find out which kernel is the best for classification on the passed bills such as linear, radial basis function (RBF), sigmoid, and polynomial using cross validation with the value of K equals to 10. The evaluation shows that the model reaches the highest F1-score using radial basis function kernel, C=1 and gamma=1 with 96,36% score.

Keywords: kernel function, law on the elimination of sexual violence, sentiment analysis, support vector machine.

Introduction

The high rate of sexual violence in Indonesia has raised concerns from the National Commission on Violence Against Women (Komnas Perempuan). In 2021 there were 338,496 complaints of cases of gender-based violence based on complaints to Komnas Perempuan, service agencies, or the Religious Courts. A total of 4,660 out of 16,162 or 28.8% of the complaints received by Komnas Perempuan and service agencies were cases of sexual violence (Komnas Perempuan, 2022). Based on this data, Komnas Perempuan initiated the making of the Draft Law on Criminal Sexual Violence (RUU PKS) since 2012 and then in May 2016 the RUU PKS was discussed by the DPR RI. The long journey in the National Legislation Program (Prolegnas) from 2016 resulted on April 12th, 2022 where the RUU PKS was passed into the law on the elimination of sexual violence (UU TPKS) (Guzman, 2022).

This ratification created various reactions from the Indonesian people both on the pros and cons. Some Indonesian women feel that the passage is a good first step towards securing women's lives and sense of security in a day-to-day life for both men and women, but some are skeptical of the law's effectiveness. Various opinions and euphoria from the ratification of the TPKS Law were poured into social media, one of which was Twitter. Based on Dixon (2022), Indonesia is the country with the fifth highest number of Twitter users. Users in Indonesia reach 18.45 million people, making Twitter a potential place to mine data for sentiment analysis.

Sentiment analysis is a field of study to analyze opinions, sentiments, evaluations, behavior, and emotions of a person based on written language. Sentiment analysis systems are used in almost every business and social area because opinions are central to almost all human activities and a key influence on personal behavior (Liu, 2012). The main purpose of sentiment analysis is to classify the author's behavior towards certain topics into three categories, namely positive, negative, and

ISSN 2829-3355 | EISSN 2828-8467

natural (Beigi et al, 2016). Sentiment analysis can be called opinion mining because it shows the same field of study and can be considered in the sub-area of subjectivity analysis (Pang and Lee, 2008).

This study aims to compare the function of the kernel in classifying public sentiment regarding the ratification of the UU TPKS. The result is the best kernel function with the combination of C, gamma, and/or degree which can then be used in future research.

Ihsan et al (2021) analyzes the sentiments of the Indonesian people towards the draft law on the law and criminal code (RKUHP), which is considered over-criminalization. This research uses the SVM algorithm with the crowdsourcing method of labeling and weighting using the TF-IDF and sentiment dictionary. The results of the evaluation of the model using cross validation with a value of K = 10 accompanied by a mean approach produces an accuracy value of 95% using the kernel radial basis function, C = 1000 and gamma = 0.0001.

Sentiment analysis research for Indonesian-language texts using the Maximum Entropy and SVM algorithms respectively produced an accuracy of 86.81% on the 7-fold cross validation test of the Sigmoid kernel function and manual labeling with POS tagger resulted in an accuracy of 81.67%. The study used the POS tagger to produce a classification model through the training process and TF-IDF as a weighting algorithm (Putranti & Winarko, 2014).

Materials and Methods

Study area

1. *Pre-processing*

   The dataset used in this study was collected from Twitter using scraping technique with "UU TPKS" and "#UU TPKS" as keywords from April 12 to April 20, 2022. The scraping process is carried out at Google Collaboratory and uses the social networking services (SNS) package. The data that has been downloaded will be sorted so that you get data that does not deviate from the topic to be discussed. Next is labeling the raw data for each tweet in the form of sentences with positive (1) and negative (0) sentiment.

   There are a few stages in the pre-processing. The stages are as follows:

   i. Drop irrelevant tweets and null.
   ii. Cleansing: remove characters that have no effect and only become noise in the data.
   iii. Case-folding: converts all letters into lowercase and eliminates repeated letters.
   iv. Stopword removal: remove words that have no meaning or stop words and save important words.
   v. Stemming: change all the words with affixes into its root words.

2. *Term Frequency and Inverse Document Frequency (TF-IDF)*

   Two different words combined into one i. e. Term Frequency and Inverse Document Frequency, form TF-IDF. TF is used to measure how many times a term is present in a document. The equation of TF is as follows

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}, \quad (1)$$

   where t is the resulting term, d is the document, $f_{t,d}$ is the number of t in d, and t' is any other number besides t.

   The inverse document frequency the opposite of TF, assigns lower weight to frequent words and assigns greater weight for the words that are infrequent (Qaiser & Ali, 2018). The equation of IDF is as follows

$$idf(t,D) = \log \frac{N}{1+|\{d \in D, t \in d\}|}, \quad (2)$$

   where N is how many documents are in the corpus and the value $N = |D|$.

3. *Support Vector Machine (SVM)*

   In this study, we used SVM with different kernel functions for classification of public sentiment. There are many algorithms being used to classify sentiment analysis. SVM is known as the most promising on classifying text. In general, the two separators are called hyperplanes which have the same function as

dividing the sample into two. The hyperplane that separates the samples has a maximum margin distance (maximum-margin hyperplane) calculated from the hyperplane to the nearest vector (Berry & Kogan, 2010).

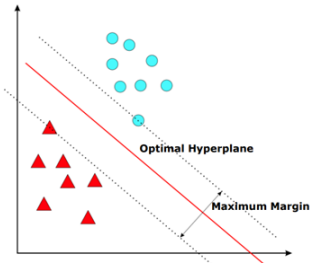

Figure 2. *The maximum margin separating the two classes (Hussain et al, 2011)*

However, data cannot always be separated using a linear equation, it is possible that a data has dimensions that cannot be separated linearly. Therefore, you can use kernel functions as a solution to provide additional dimensions to the hyperplane. The kernel function is a mathematical trick that allows SVM to classify dimensional spaces that have more dimensions than the data it has. For example, if data has one dimension, the kernel runs SVM to perform classification on two or more of these data (Phienthrakul et al, 2009).

There are many kernel functions and in this study we used four kernel functions to compare its performance. The four kernels and its equation are as follows:

i. Linear kernel with cost C can be written as

$$K(x, y) = x.y. \quad (3)$$

ii. Radial basis function (RBF) kernel with cost C and gamma can be written as

$$K(x, y) = \exp(-\gamma\|x - y\|^2). \quad (4)$$

iii. Sigmoid kernel with cost C and gamma is given as

$$K(x, y) = \tanh(x.y + C). \quad (5)$$

iv. Polynomial kernel with cost C, gamma, and degree d is given as

$$K(x, y) = (x.y + C)^d. \quad (6)$$

4. *F1-score*

There are many evaluation metrics that can be used to evaluate the performance of a classifier. As for this research, we used F1-Score as an evaluation metric to evaluate the trained classifier and find the best way to classify data testing and future data. This evaluation is a great measure because it favours algorithms with higher sensitivity and challenges those with higher specificity (Sokolova et al, 2006). F1-Score is a metric that represents the harmonic mean between recall (r) and precision (p) values which are also evaluation metrics. These metrics can be computed from the confusion matrix as shown in Table 1. The row part of the table represents the predicted class from the algorithm, while the column represents the actual class.

Table 1. Confusion matrix.

	Predicted Positive Class	Predicted Negative Class
Actual Positive Class	True Positive (tp)	False negative (fn)
Actual Negative Class	False positive (fp)	True negative (tn)

p is a metric that is used to measure the positive patterns that the algorithm correctly predicted from the total predicted patterns in a positive class. The formula of p is as shown below.

$$p = \frac{tp}{tp+tn}. \quad (7)$$

r is a metric that is used to measure the fraction of positive patterns that are correctly classified using the formula shown below.

$$r = \frac{tp}{tp+fp}. \quad (8)$$

From those metrics, we can measure the F1-score which formula shown below.

$$F1 - score = \frac{2*p*r}{p+r}. \quad (9)$$

F1-score is a great discriminator and performed better than other basic metrics for binary classification problems (Hossin & Sulaiman, 2015)

Procedures

In this research, we used data that has been scrapped from twitter using keywords "UU TPKS" and "#UUTPKS". The data are limited from April 12th, 2022 the day UU TPKS passed into law until April 20th, 2022. In this study, we used 5,486 tweets from those keywords and the method used is SVM with four kernel functions including linear, radial basis function, sigmoid and polynomial. The stages below are carried out in this research as shown.

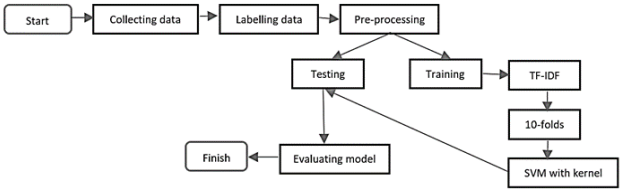

Figure 2. *Methodology flowchart.*

i. Collecting data on Twitter using keywords "UU TPKS" or "#UUTPKS".
ii. Labeling the raw data for each tweet in the form of sentences with positive (1) and negative (0) sentiment.
iii. Preprocessing data consists dropping irrelevant tweets, cleansing, case-folding, stop-word removal, stemming, and tokenizing.
iv. Weighting with Term Frequency-Inverse Document Frequency (TF-IDF).
v. Splitting data into training and testing data.
vi. Classifying using SVM with kernel function and the combination of C, gamma, and degree.
vii. Evaluating model using F1-score.

Results and Discussion

The data scraping process is carried out at the Google colaboratory using the Python 3 programming language with the SNScrape library. By using this library, we can scrape data from Twitter without Twitter API without limitation, whereas Twitter API can only retrieve data for the past week. Using SNScrape library, we obtained 6,336 tweets using "UU TPKS" and "#UUTPKS" as keywords. The duplicated and irrelevant tweets are removed and we have 5,486 raw data. Next, we labelled all of the raw data with positive and negative sentiment, which 5,091 data are labelled positive and 395 data are labelled negative.

The raw data from Twitter by using scraping are tweets that are not well-structured. Before stepping into the classification, we used several pre-processing steps to make the raw data into well-structured data. The example of the pre-processing result as shown on Table 2.

Table 2. An example of the preprocessing result.

Stage	Tweet
Raw Tweet	@Adiiiinnnnnn Harusnya adik2 juga mengapresiasi disahkannya UUTPKS, jangan malah pamer aset berhargamu pada begal kekerasan s3ksu4l
Cleansing	Adin Harusnya adik juga mengapresiasi disahkanya UTPKS jangan malah pamer aset berhargamu pada begal kekerasan sksul
Case-folding	adin harusnya adik juga mengapresiasi disahkanya utpks jangan malah pamer aset berhargamu pada begal kekerasan sksul
Stopword removal	adin harusnya adik mengapresiasi disahkanya utpks jangan malah pamer aset berhargamu begal kekerasan sksul
Stemming	adin harus adik apresiasi disahkanya utpks jangan malah pamer aset harga begal keras sksul
Tokenizing	['adin', 'harus', 'adik', 'apresiasi', 'disahkanya', 'utpks', 'jangan', 'malah', 'pamer', 'aset', 'harga', 'begal', 'keras', 'sksul']

After pre-processing data, the next step is weighting all the terms on the tweets using TF-IDF to measure how frequently a term appears in a document and how important a term is. TF-IDF evaluates how relevant a word is to a document in a collection of documents and improve the value of recall and precision.

In this research, we used 10-fold Cross-validation as a validator and confusion matrix as evaluator. F1-score is used as an evaluation metric and we divided tweets into 80% as a training data and 20% as testing data. Based on Table 3, the SVM algorithm using radial basis function with $C = 1$ and $\gamma = 1$ outperformed other kernel functions. Some of the result come out with similar F1-score, means those has the same performance on the classification. All of kernels has performances above 95%.

Table 3. F1-Score of kernel functions.

Algorithm	Cost (C)	Gamma (γ)	Degree (d)	F1-Score
Linear	1	-	-	96,16%
	10	-	-	95,28%
RBF	**1**	**1**	**-**	**96,36%**
	1	0.5	-	96,26%
	10	1	-	96,24%
	10	0.5	-	96,16%
Sigmoid	1	1	-	96,16%
	1	0.5	-	96,26%
	10	1	-	95,32%
	10	0.5	-	95,63%
Polynomial	1	1	2	96,26%
	1	0.5	2	96,27%
	1	1	3	96,26%
	1	0.5	3	96,27%
	10	1	2	96,21%
	10	0.5	2	96,21%
	10	1	3	96,16%
	10	0.5	3	96,21%

Conclusions

Based on the discussion, the result for classification on the sentiment analysis of the law on the elimination of sexual violence using SVM kernel functions using 10-fold Cross-validation and TF-IDF shows that the radial basis function (RBF) kernel with the combination of $C = 1$ and $\gamma = 1$ has achieve the best classification with 96.36% F1-score. Meanwhile, all other kernels also give their best classification which are all above 95% F1-score.

References

Beigi, G., Hu, X., Maciejewski, R., Liu, H. 2016. An Overview of Sentiment Analysis in Social Media and Its Applications in Disaster Relief. In: Pedrycz, W., Chen, SM. (eds) Sentiment Analysis and Ontology Engineering. Studies in Computational Intelligence, vol 639. Springer, Cham. https://doi.org/10.1007/978-3-319-30319-2_13

Berry, M. W., & Kogan, J. (Eds.). 2010. Text mining: applications and theory. John Wiley & Sons.

Dixon, S. 2022. Countries with most Twitter users 2022. Statista. https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

Guzman, C. de. 2022. Indonesia Finally Passes Sexual Violence Bill. Time. https://time.com/6166853/indonesia-sexual-violence-law/

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process, 5(2), 1.

Hussain, M., Wajid, S. K., Elzaart, A., & Berbar, M. 2011. A Comparison of SVM Kernel Functions for Breast Cancer Detection. 2011 Eighth International Conference Computer Graphics, Imaging and Visualization. doi:10.1109/CGIV.2011.31

Ihsan, I., Nurjanah, D., & Nurrahmi, H. 2021. Sentiment Analysis RKUHP Pada Twitter Menggunakan Metode Support Vector Machine. eProceedings of Engineering, 8(2).

Komnas Perempuan. (2022). Bayang-Bayang Stagnansi: Daya Pencegahan dan Penanganan Berbanding Peningkatan Jumlah, Ragam dan Kompleksitas Kekerasan Berbasis Gender terhadap Perempuan. Jakarta: Komnas Perempuan.

Liu, B. 2012. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

Pang, B., & Lee, L. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1–2), 1–135.

Phienthrakul, T., Kijsirikul, B., Takamura, H., & Okumura, M. 2009. Sentiment Classification with Support Vector Machines and Multiple Kernel Functions. Lecture Notes in Computer Science, 583–592. doi:10.1007/978-3-642-10684-2_65

Prastyo, P. H., Ardiyanto, I., & Hidayat, R. (2020). Indonesian Sentiment Analysis: An Experimental Study of Four Kernel Functions on SVM Algorithm with TF-IDF. 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI). doi:10.1109/ICDABI51230.2020.9325685

Qaiser, S., & Ali, R. 2018. Text mining: use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications*, 181(1), 25-29.

Qaiser, S., & Ali, R. 2018. Text mining: use of TF-IDF to examine the relevance of words to documents. International Journal of Computer Applications, 181(1), 25-29.

Rozi, I. F., Pramono, S. H., & Dahlan, E. A. 2013. Implementasi opinion mining (analisis sentimen) untuk ekstraksi data opini publik pada perguruan tinggi. *Jurnal EECCIS*, 6(1), 37-43.

Sasaki, Y. (2007). The truth of the F-measure. Teach tutor mater, 1(5), 1-5.

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006, December). Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In Australasian joint conference on artificial intelligence (pp. 1015-1021). Springer, Berlin, Heidelberg.

Wijaya, T. N., Indriati, R., & Muzaki, M. N. 2021. Analisis Sentimen Opini Publik Tentang Undang-Undang Cipta Kerja Pada Twitter. *Jambura Journal of Electrical and Electronics Engineering*, 3(2), 78-83.