Recursive Journal of Informatics
Vol. 2, No. 1, September 2023
p-ISSN 2963-5551 https://journal.unnes.ac.id/sju/index.php/rji e-ISSN 2986-6588

Sentiment Analysist of the TPKS Law on Twitter Using InSet Lexicon with Multinomial Naïve Bayes and Support Vector Machine Based on Soft Voting

Salsabila Rahadatul Aisy[1], Budi Prasetiyo[2]
[1,2]Computer Science Department, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Semarang, Indonesia

Abstract. The Indonesian Sexual Violence Law (TPKS Law) is a law that regulates forms of sexual violence. The TPKS Law reaped pros and cons in the drafting process and was officially ratified on April 12th, 2022. However, after being ratified, pros and cons can still be found and supervision is needed over the implementation of the law.
Purpose: This study was conducted to identify the application and accuracy of soft voting on multinomial naïve Bayes and support vector machine algorithm, also to find out public opinion on the TPKS Law as a support tool in evaluating the law.
Methods/Study design/approach: The method used is InSet lexicon for labeling with the soft voting classification method on the multinomial naive Bayes and support vector machine algorithm.
Result/Findings: The accuracy obtained by applying 10 k-fold cross validation in soft voting is 84.31%, which uses a weight of 1:3 for multinomial naive Bayes and support vector machines. Soft voting obtains better accuracy than its standalone predictor, and also works well for sentiment analysis of the TPKS Law.
Novelty/Originality/Value: This study using two combined lexicons (Colloquial Indonesian lexicon and the InaNLP formalization dictionary) in normalization process and using InSet lexicon as automatic labeling for sentiment analysis on TPKS Law.

Keywords: sentiment analysis, TPKS law, multinomial naïve Bayes, support vector machine, soft voting, InSet lexicon
Received May 05, 2023 / **Revised** May 31, 2023 / **Accepted** September 14, 2023

INTRODUCTION

Sentiment analysis is a tool capable of extracting text that contains subjective information about an opinion or sentiment [1]. Sentiment analysis is mostly done on social media to analyze the sentiment of certain topics. One of the social media that is widely used as information to understand public opinion is Twitter [2]. Information available on social media can be used to identify knowledge in providing better products and services [3]. There are three methods that can be used in sentiment analysis, namely machine learning, lexicon-based, and hybrid methods [4].

Machine learning is a branch of artificial intelligence which has the goal of analyzing data, then learning from the results obtained to make predictions regarding data that being analyzed [5]. Training process in machine learning requires labelled data, therefore dataset labeling is needed. The labeling process can be done using a lexicon, which can reduce the time and cost of manual labeling [6] on large amounts of data. One lexicon that can be used to determine sentiment as a label is the Indonesian Sentiment (InSet) lexicon. The InSet lexicon is an Indonesian lexicon that consisting of 6,609 negative words and 3,609 positive words with a score between -5 to +5, and has better performance than several other lexicons such as Vanilla lexicon, SentiWordNet, Liu lexicon, or Afinn lexicon [7].

Dataset that already has label can be trained and tested using machine learning algorithms. In sentiment analysis, naïve Bayes and support vector machines are popular and widely used machine learning algorithms [8]. Naïve Bayes is known as a simple powerful model, especially in the field of document classification and disease prediction [9], while the support vector machine is a machine learning algorithm

[1]*Corresponding author.
Email addresses: salsabilarhdsy@students.unnes.ac.id (Aisy)
DOI: 10.15294/rji.v1i2.68324

that has a good and efficient training stage, hence good performance is obtained [10]. Previous study using multinomial naïve Bayes and support vector machines has been done by [11]. The study discusses the classification of anti-Islamic texts in Arabic by comparing several algorithms, such as k-NN, decision tree, random forest, logistic regression, multinomial naïve Bayes and support vector machine. In that study, the naïve Bayes multinomial algorithm and support vector machine obtained better accuracy than other algorithms.

Sentiment analysis using multinomial naïve Bayes and support vector machines has been done in many previous studies, but rarely discussed regarding the Indonesia sexual violence law (TPKS Law). The TPKS Law is a law that regulates forms of sexual violence and guarantees protection for the victims [12]. During the discussion, the draft law reaped pros and cons among the public. The difference is triggered by a person's perspective of seeing a rule, which refers to the perspective of gender and religious morality [13]. The ratified of the TPKS Law was a breakthrough in law enforcement for cases of sexual violence [12]. However, pros and cons are still being found after it was officially ratified [14] and there is still a need supervision for the implementation of the TPKS Law [15]. Therefore, sentiment analysis is needed to find out how public opinion is regarding the TPKS Law as a tool support in evaluating the law.

Previous study using multinomial naïve Bayes algorithms and support vectors was also done by [16], which discusses sentiment analysis regarding extremism on social media. In the study, preprocessing was done including data cleaning and case folding. The accuracy results obtained were 66% for multinomial naïve Bayes and 82.1% for support vector machines. Besides using multinomial naïve Bayes and support vector machines as standalone predictors for classification, the performance of standalone predictors can be improved using the ensemble method [17].

The ensemble method is a method that combines different data mining techniques to find predictions [18]. One of the ensemble methods is the classifier voting method, which is divided into hard voting and soft voting. Hard voting is a method that combines several machine learning models where the prediction result is the result of the most votes of all the models used [19]. Besides, soft voting is a method that predicts by calculating the average probability of each class of all models used, then taking the highest value [20]. In previous study, this method was able to obtain better accuracy than the standalone predictor method [21] [17], or compared to the hard voting method [21].

Based on the above description, this study focuses on analyze the sentiments of the TPKS Law using InSet lexicon as an automatic labeling with multinomial naive bayes and support vector machine algorithm based on soft voting. The purpose of this study is to identify the application and accuracy of soft voting method.

METHODS

The study was conducted using InSet lexicon as automatic labeling with classification using multinomial naïve Bayes and support vector machine based on voting. Flowchart of sentiment analysis process conducted is addressed in Figure 1. The study started with collecting data from Twitter using Snscrape library.

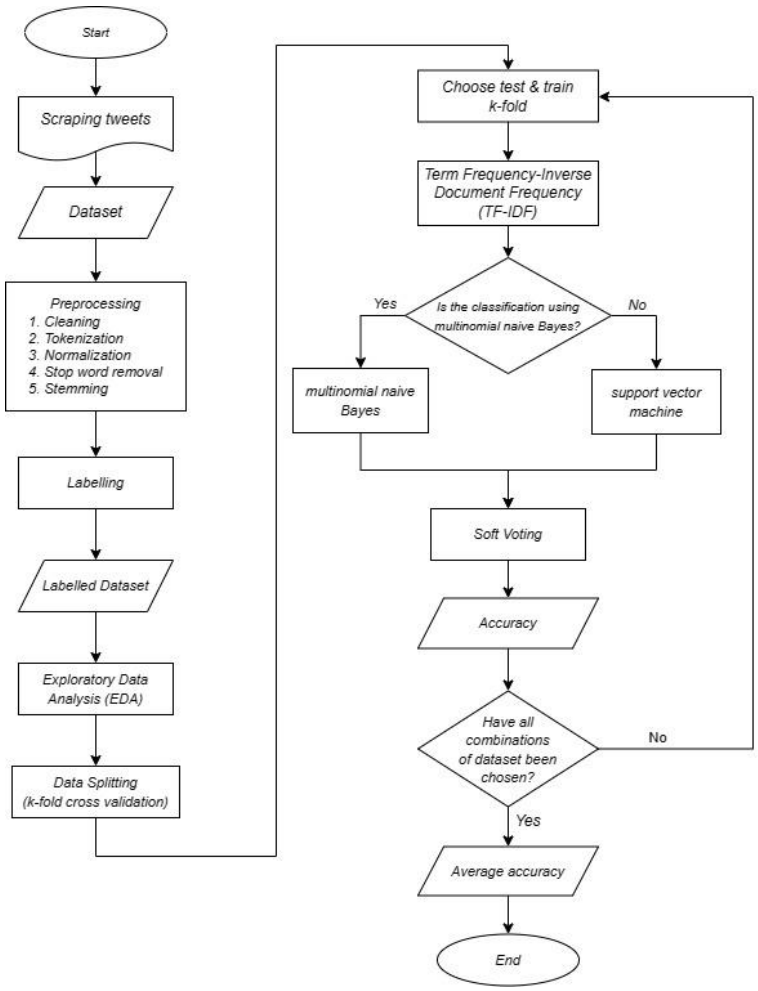

Figure 1. Flowchart Research Design

Preprocessing Data

In the text mining process, unstructured text requires preprocessing to prepare the text before going to the next stage to make it more structured [22]. Preprocessing carried out includes data cleaning, tokenization, normalization, stop words removal, and stemming. Data cleaning includes case folding, which changes all text to lowercase, removing retweet tags, usernames, URLs, hashtags, numbers, and special characters, deleting letters in a word that repeats more than 2 times, for example the word "naaah" becomes "naah", also deleting duplicate tweets. Afterwards, normalization was carried out using a combination of the Colloquial Indonesian lexicon and the InaNLP formalization dictionary.

Data Labelling

Lexicon is a method used in sentiment analysis for labelling. This method works with matches every word in a dataset into the lexicon to find polarity, which includes positive or negative sentiments [23]. In its application, data tweets that have been preprocessed will be matched into the InSet lexicon. The total polarity value is obtained from the accumulation of word weights on negative and positive lexicon matches. Each tweet labeled as its polarity, either positive (polarity > 0), neutral (polarity = 0) or negative (polarity < 0). The results of the labeling process are visualized at the exploratory data analysis stage using diagrams and word cloud.

Cross Validation

Cross validation is a method for evaluating machine learning algorithms where the dataset is divided into k parts and evaluates each sample to get average accuracy [24]. In the process, the training data will be divided into k parts, then the i^{th} data is taken as testing data and the rest is training data [25]. The method used in this process is 10 k-fold cross validation, where this method tends to provide accurate results with

minimal bias [26]. By using 10 k-fold cross validation, the comparison of training data and test data is 90:10.

Multinomial Naïve Bayes

Naïve Bayes is a machine learning algorithm used to classify text such as sentiment analysis, which is able to create and predict models quickly, and can have excellent results for text data analysis [9]. Naïve Bayes as a machine learning algorithm that uses a probabilistic approach tends to work well for handling training sets that change over time [27]. One type of naïve Bayes is multinomial naïve Bayes. This method uses probability and focuses on text classification cases, and follows the principle of multinomial distribution [28].

Support Vector Machine

Support vector machine is a machine learning algorithm that classifies binary classes by creating an N-dimensional hyperplane that divides the two classes optimally [29]. Support vector machine that used for classification is also called Super Vector Classification (SVC). SVC can form a hyperlane with the maximum distance (margin) from the support vector in linear data [10].

Soft Voting

Soft voting is a type of voting classifier. The voting method can combine the results of two or more classification models [20]. The soft voting classifier works by calculating the average probability for each class of all the models used, then taking the highest value [20]. In soft voting, predictions are weighted based on how important the classifier is, then the class with the highest weighted probabilities value will be selected by calculating the average [30].

Confusion Matrix

The confusion matrix is a matrix that contains classification data carried out by a classification model, both actual data and predictive data, thus the performance of the model can be calculated [31]. In the study of [31] explained the confusion matrix using two classes, that the positive and negative classes, which shown in Table 1 and obtained an equation to calculate accuracy in Equation 1.

Table 1. Confusion matrix

		Prediction	
		Positive	Negative
Actual	Positive	TP (True Positive)	FN (False Negative)
	Negative	FP (False Positive)	TN (True Negative)

$$\text{accuracy} = \frac{TP+TN}{TP+FP+FP+FN}\ x\ 100\% \quad (1)$$

RESULT AND DISCUSSION

The study aims to identify application and to know the results of soft voting performance in predicting sentiment of TPKS Law tweets. The study was done in several stages, including data scraping, preprocessing, data labeling using InSet lexicon, exploratory data analysis, data splitting, TF-IDF, and classification using multinomial naïve Bayes and support vector machine based on soft voting.

Collecting data in this study was carried out using Snscrape library with keyword ‘uu tpks’ from 12 to 18 April 2022. The amount of data obtained was 6,872 tweets. Following by preprocessing, which data cleaning, tokenization, normalization, stop words removal and stemming. In the data cleaning stage, duplicate tweets are also removed from the dataset, so that the amount of data is reduced from 6,872 tweets to 6,405 tweets. Afterward, at the tokenization stage, the data is converted into words or tokens. Furthermore, normalization is carried out to change slang words into standard or formal words, for example the word 'udah' becomes 'sudah', 'ngebuktiin' becomes 'membuktikan'. The normalization stage was carried out using a combination of the colloquial Indonesian lexicon and the formalization dictionary from InaNLP. The two lexicons combined have 5,079 words.

The stop words removal and stemming stages were carried out using the Sastrawi library. Example of Sastrawi stop words are ‘yang’, ‘untuk’, and ‘pada’. Other stop words were also added, such as the uu tpks keyword, namely the words 'uu', 'tpks', and 'uutpks'. Then add a few words that are included in the negative InSet lexicon but don't have a negative meaning in the context of opinion tweets against uu tpks, such as

the words 'victim', 'crime', and other words that don't have opinion meanings such as ‘pahlawan’, ‘puan’, ‘indonesia’. Figure 2 shows the words added to the stop words data. While the results of the preprocessing stages can be seen in Table 2.

'uu', 'ruu', 'rancangan', 'ruutpks', 'undang', 'tpks', 'uutpks', 'tindak', 'pidana', 'kekerasan', 'seksual', 'korban', 'kejahatan', 'pelecehan', 'pemaksaan', 'rapat', 'penjara', 'diancam', 'ancaman', 'pasal', 'implementasi', 'diimplementasikan', 'mengimplementasikan', 'aborsi', 'pahlawan', 'puan', 'maharani', 'dpr', 'indonesia', 'perempuan', 'memperjuangkan', 'diperjuangkan', 'perjuangan', 'perjuangannya', 'ri', 'berikut', 'atas', 'sih', 'apa-apa', 'amp', 'aa', 'iya', 'si', 'eh', 'kak', 'oh', 'he', 'nder'

Figure 2. Extend stop words list

Table 2. Result of preprocessing

Tweet	Data cleaning	Tokenization	Normalization	Stop words removal	Stemming
Puan Maharani Akan Pimpin Rapat Paripurna RUU TPKS Menjadi UU https://t.co/gLz7HM4iNC	puan maharani akan pimpin rapat paripurna ruu tpks menjadi uu	[puan, maharani, akan, pimpin, rapat, paripurna, ruu, tpks, menjadi, uu]	[puan, maharani, akan, pimpin, rapat, paripurna, ruu, tpks, menjadi, uu]	[pimpin, paripurna]	[pimpin, paripurna]
@andianisharfina @maidina__ Udah bisa pake nama #UUTPKS udah sah	udah bisa pake nama udah sah	[udah, bisa, pake, nama, udah, sah]	[sudah, bisa, pakai, nama-nama, sudah, sah]	[pakai, nama-nama, sah]	[pakai, nama, sah]
…	…	..	…	…	…
@rizkiwulandani @byeoreum @tegarfr__ @dikignwn @Wirogendheng88 @tubirfess UU TPKS dan kenaikan usia menikah aja udah ngebuktiin ini salah, child marriage itu dihalalkan di islam. Sedangkan batas usia pernikahan itu 19 disini	uu tpks dan kenaikan usia menikah aja udah ngebuktiin ini salah *child marriage* itu dihalalkan di islam sedangkan batas usia pernikahan itu disini	[uu, tpks, dan, kenaikan, usia, menikah, aja, udah, ngebuktiin, ini, salah, child, marriage, itu, dihalalkan, di, islam, sedangkan, batas, usia, pernikahan, itu, disini]	[uu, tpks, dan, kenaikan, usia, menikah, saja, sudah, membuktikan, ini, salah, child, marriage, itu, dihalalkan, di, islam, sedangkan, batas, usia, pernikahan, itu, disini]	[kenaikan, usia, menikah, membuktikan, salah, child, marriage, dihalalkan, islam, batas, usia, pernikahan]	[naik, usia, meni, bukti, salah, child, marriage, halal, islam, batas, usia, nikah]
UU TPKS Disahkan, Berikut 6 Poin Penting dari Permendikbudrisek https://t.co/rtxsuhD4ox	uu tpks disahkan berikut poin penting dari permendikbudristek	[uu, tpks, disahkan, berikut, poin, penting, dari, permendikbudristek]	[uu, tpks, disahkan, berikut, poin, penting, dari, permendikbudristek]	[disahkan, poin, permendikbudristek]	[sah, poin, permendikbudristek]

The results of the InSet lexicon labeling process on the TPKS dataset can be seen in Table 3. The score results from matching words to the negative lexicon are shown in the negative score column, as well as the results from matching words to the positive lexicon which can be seen in the positive score column. Both scores are accumulated and stored in the total score column. The total score or polarity determines the labeling result, either positive, neutral or negative.

Table 3. Labelled Dataset

Teks	Negative score	Positive score	Total score	Label
[pimpin, paripurna]	0	1	1	Positive
[pakai, nama, sah]	0	0	0	Neutral
…	…	…	…	…
[naik, usia, meni, bukti, salah, child, marriage, halal, islam, batas, usia, nikah]	-15	11	-4	Negative
[sah, poin, permendikbudristek]	0	2	2	Positive

The results of the 6,405 data used, the most data obtained on April 12, 2022 with 2,745 data, and the least data obtained on April 17, 2022 with 142 data. This is shown in Figure 3. While for each sentiment, there are 3,056 data labeled positive, 1,188 data labeled neutral, and 2,161 data labeled negative as shown in Figure 4.

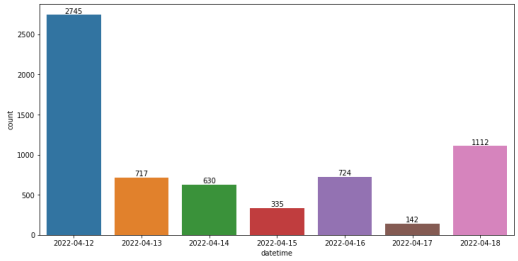

Figure 4. Data amount by date

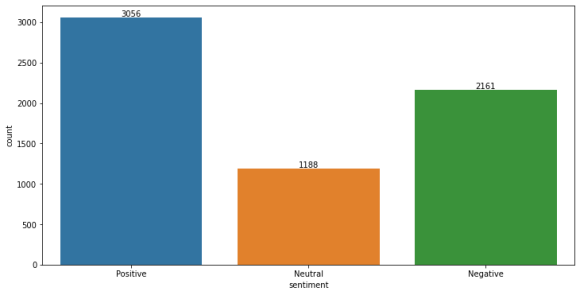

Figure 3. Data amount by sentiment

In 10 k-fold cross validation, the data is divided into 10 parts and multiplied 10 times. With 10 k-fold cross validation, 90% of the training data was obtained and 10% of the test data with different combinations for each iteration. Then each iteration goes through the TF-IDF and classification stages. The accuracy obtained in the classification process is the result of the confusion matrix calculation on data testing. To obtain the parameters with the best accuracy, experiments were carried out on each method. As in multinomial naïve Bayes, 12 experiments were carried out on the value of the alpha parameter and the value alpha = 0.01 resulted in the best accuracy of 74.30%. The support vector machine was carried out 13 times on parameter C and the best accuracy was obtained is 83.93% on parameter C=4, also soft voting was carried out on experiments of 22 different weights obtain the best accuracy in 84.31% for a weight of 1:3. The accuracy results for each iteration of each method are shown in Table 4.

Table 4. MNB, SVM, and soft voting accuracy results

Iteration	Multinomial naïve Bayes	Support vector machine	Soft voting
1	68.33%	82.21%	82.68%
2	73.47%	83.15%	82.83%
3	83.61%	91.88%	91.73%
4	75.19%	85.17%	86.11%
5	72.54%	80.03%	80.49%
6	71.25%	80.15%	80.93%
7	70.62%	78.12%	78.75%
8	76.56%	87.50%	86.87%
9	75.31%	85%	85.62%
10	76.09%	86.09%	87.03%
average	**74.30%**	**83.93%**	**84.31%**

Discussion

Compared to previous studies in Table 5, the accuracy of this study was 84.31%, which was 2.61% to 14.2% higher. Based on the accuracy results obtained, the soft voting method can work well in the sentiment analysis of the TPKS Law. Furthermore, when compared to previous studies using manual labeling [21], [32], [33], [34], this study was able to obtain better accuracy using automatic labeling, which is the InSet lexicon. [35] in their study concluded that using the lexicon for automatic labeling is able to obtain better accuracy compared to manual labeling.

Furthermore, research by [16] who have similarities in the automatic labeling process obtain 82.1% accuracy on support vector machines and 66% on multinomial naïve Bayes. The difference between that study and this study is the preprocessing that was carried out. In the study by [16], only data cleaning is used for preprocessing. The preprocessing stages in this study are data cleaning, tokenization, normalization, stop words removal, and stemming. According to [36], the right preprocessing technique plays an important role in cleaning the data and increasing the accuracy of the classifier used. Therefore, this research can obtain better accuracy by using other preprocessing stages, such as normalization, stop words removal, and stemming.

Table 5. Accuracy comparison with previous studies

Authors	Data	Language	Method	Accuracy
Delizo et al. (2020)	Twitter	Tagalog	Multinomial naïve Bayes	72.00%
Asif et al. (2020)	Facebook	Multilingual	Multinomial naïve Bayes Support vector machine	66.00% 82.10%
Khurniawan & Ruldeviyani, (2020)	Twitter	Indonesia	Support vector machine Naïve Bayes Decision tree	81.70% 80.90% 74.55%
Sontayasara et al. (2021)	Twitter	English	Support vector machine	71.00%
Andrian et al. (2022)	Twitter	Indonesia	Hard voting Soft voting	75.43% 75.86%
Proposed method	Twitter	Indonesia	Soft voting	84.31%

CONCLUSION

Based on the results and discussion of the study that has been done, it can be concluded that the labelling process using the InSet lexicon is carried out by matching each word in the dataset to the positive and negative InSet lexicon. Tweets that have a total weight or polarity> 0, then the tweet is labelled positive, if polarity = 0 then it is labelled neutral, and if polarity <0 then it is labelled negative. Furthermore, an analysis of public sentiment towards the TPKS Law was carried out through several stages. These stages include the process of collecting data through scraping, then the preprocessing stage which includes data cleaning, normalization, stop words removal, and stemming. Afterwards, the labelling stage was carried out using the InSet lexicon, data visualization using EDA, the data splitting stage using 10 k-fold cross validation, and feature extraction using TF-IDF. After those stages, the data will be trained and tested using the multinomial naïve Bayes method, support vector machine, and soft voting in order to obtain model accuracy. Final conclusion, the accuracy results obtained using multinomial naïve Bayes is 74.30%, support vector machine is 83.93%, and soft voting is 84.31%. The best accuracy is obtained when using soft voting with a weight of 1:3 for multinomial naïve Bayes and support vector machines, which because the value of the weight parameter given to the support vector machine is higher than the weight value in multinomial naïve Bayes.

REFERENCES

[1] F. A. Pozzi, E. Fersini, E. Messina, and B. Liu, "Challenges of sentiment analysis in social networks: An overview," in *Sentiment Analysis in Social Networks*, Elsevier Inc., 2017, pp. 1–11. doi: 10.1016/B978-0-12-804412-4.00001-2.

[2] A. Reyes-Menendez, J. R. Saura, and C. Alvarez-Alonso, "Understanding #worldenvironmentday user opinions in twitter: A topic-based sentiment analysis approach," *Int J Environ Res Public Health*, vol. 15, no. 11, pp. 1–18, Nov. 2018, doi: 10.3390/ijerph15112537.

[3] A. F. Abbas, A. Jusoh, A. Mas'od, A. H. Alsharif, and J. Ali, "Bibliometrix analysis of information sharing in social media," *Cogent Business & Management*, vol. 9, no. 1, pp. 1–23, 2022, doi: 10.1080/23311975.2021.2016556.

[4] A. Sadia, F. Khan, and F. Bashir, “An overview of lexicon-based approach for sentiment analysis,” in *3rd International Electrical Engineering Conference (IEEC)*, 2018, pp. 1–6.
[5] R. Štrimaitis, P. Stefanovič, S. Ramanauskaitė, and A. Slotkienė, “Financial context news sentiment analysis for the Lithuanian language,” *Applied Sciences (Switzerland)*, vol. 11, no. 10, pp. 1–13, May 2021, doi: 10.3390/app11104443.
[6] S. Biswas, K. Young, and J. Griffith, “A comparison of automatic labelling approaches for sentiment analysis,” *arXiv:2211.02976v1*. Nov. 05, 2022. doi: 10.5220/0011265900003269.
[7] F. Koto and G. Y. Rahmaningtyas, “Inset lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs,” in *Proceedings of the 2017 International Conference on Asian Language Processing, IALP 2017*, Institute of Electrical and Electronics Engineers Inc., Feb. 2018, pp. 391–394. doi: 10.1109/IALP.2017.8300625.
[8] D. A. Kristiyanti, D. A. Putri, E. Indrayuni, A. Nurhadi, and A. H. Umam, “E-wallet sentiment analysis using naïve Bayes and support vector machine algorithm,” *J Phys Conf Ser*, vol. 1641, no. 1, Nov. 2020, doi: 10.1088/1742-6596/1641/1/012079.
[9] M. Abbas, K. Ali Memon, A. Aleem Jamali, S. Memon, and A. Ahmed, “Multinomial naive Bayes classification model for sentiment analysis,” *IJCSNS International Journal of Computer Science and Network Security*, vol. 19, no. 3, pp. 62–67, 2019.
[10] R. Gholami and N. Fakhari, “Support vector machine: Principles, parameters, and applications,” in *Handbook of Neural Computation*, Elsevier, 2017, pp. 515–535. doi: 10.1016/B978-0-12-811318-9.00027-2.
[11] R. A. Alraddadi and M. I. E.-K. Ghembaza, “Anti-islamic Arabic text categorization using text mining and sentiment analysis techniques,” *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 12, no. 8, pp. 776–785, 2021, doi: 10.14569/IJACSA.2021.0120889.
[12] B. B. Bayusuta and Y. Suwanto, “Analisis yuridis undang-undang tindak pidana kekerasan seksual dalam penegakan hukum di Indonesia,” *Souvereignty : Jurnal Demokrasi dan Ketahanan Nasional*, vol. 1, no. 1, pp. 37–43, 2022.
[13] A. Caterine, B. Adi, and D. Wahyu, “Kebijakan penegakan kukum kekerasan berbasis gender online (KBGO): Studi urgensi pengesahan RUU PKS,” *Jurist-Diction*, vol. 5, no. 1, pp. 17–34, Jan. 2022, doi: 10.20473/jd.v5i1.32869.
[14] F. A. Paulina and M. Madalina, “Urgensi RUU TPKS sebagai payung hukum bagi korban kekerasan seksual beserta tantangan-tantangan dalam proses pengesahannya,” *Souvereignty: Jurnal Demokrasi dan Ketahanan Nasional*, vol. 1, no. 1, pp. 136–150, 2022.
[15] Kementerian PPPA, “Menteri PPPA ajak masyarakat kawal implementasi UU TPKS,” *Kementerian Pemberdayaan Perempuan dan Perlindungan Anak (PPPA)*, Apr. 22, 2022. https://www.kemenpppa.go.id/index.php/page/read/29/3867/menteri-pppa-ajak-masyarakat-kawal-implementasi-uu-tpks (accessed Feb. 18, 2023).
[16] M. Asif, A. Ishtiaq, H. Ahmad, H. Aljuaid, and J. Shah, “Sentiment analysis of extremism in social media from textual information,” *Telematics and Informatics*, vol. 48, no. 101345, pp. 1–20, May 2020, doi: 10.1016/j.tele.2020.101345.
[17] A. Özçift, “Medical sentiment analysis based on soft voting ensemble algorithm,” *Yönetim Bilişim Sistemleri Dergisi*, vol. 6, no. 1, pp. 42–50, 2020.
[18] A. K. Verma, S. Pal, and S. Kumar, “Classification of skin disease using ensemble data mining techniques,” *Asian Pacific Journal of Cancer Prevention*, vol. 20, no. 6, pp. 1887–1894, Jun. 2019, doi: 10.31557/APJCP.2019.20.6.1887.
[19] R. Atallah and A. Al-Mousa, “Hearth disease detection using machine learning majority voting ensemble method,” in *2nd International Conference on New Trends in Computing Sciences (ICTCS))*, IEEE, 2019, pp. 1–6.
[20] G. P. de Oliveira, A. Fonseca, and P. C. Rodrigues, “Diabetes diagnosis based on hard and soft voting classifiers combining statistical learning models,” *Brazilian Journal of Biometrics*, vol. 40, no. 4, pp. 415–427, Dec. 2022, doi: 10.28951/bjb.v40i4.605.
[21] B. Andrian, T. Simanungkalit, I. Budi, and A. F. Wicaksono, “Sentiment analysis on customer satisfaction of digital banking in Indonesia,” *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 13, no. 3, pp. 466–473, 2022, [Online]. Available: www.ijacsa.thesai.org
[22] R. Novendri, A. S. Callista, D. N. Pratama, and C. E. Puspita, “Sentiment analysis of YouTube movie trailer comments using naïve Bayes,” *Bulletin of Computer Science and Electrical Engineering*, vol. 1, no. 1, pp. 26–32, Jun. 2020, doi: 10.25008/bcsee.v1i1.5.

[23] S. Taj, B. B. Shaikh, and A. F. Meghji, "Sentiment analysis of news articles: A lexicon based Approach," in *In 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, IEEE, 2019, pp. 1–5. doi: 10.1109/ICOMET.2019.8673428.

[24] O. Karal, "Performance comparison of different kernel functions in SVM for different k value in k-fold cross-validation," in *Innovations in Intelligent Systems and Applications Conference*, IEEE, 2020, pp. 1–5.

[25] M. Niu, Y. Li, C. Wang, and K. Han, "RFAmyloid: A web server for predicting amyloid proteins," *Int J Mol Sci*, vol. 19, no. 7, pp. 1–13, Jul. 2018, doi: 10.3390/ijms19072071.

[26] P. R. Sihombing and O. P. Hendarsin, "Perbandingan metode artificial neural network (ANN) dan support vector machine (SVM) untuk klasifikasi kinerja perusahaan daerah air minum (PDAM) di Indonesia," *Jurnal Ilmu Komputer*, vol. XII, no. 1, pp. 9–20, 2020.

[27] H. A. Santoso, E. H. Rachmawanto, A. Nugraha, A. A. Nugroho, D. R. I. M. Setiadi, and R. S. Basuki, "Hoax classification and sentiment analysis of Indonesian news using naive Bayes optimization," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 18, no. 2, pp. 799–806, Apr. 2020, doi: 10.12928/TELKOMNIKA.V18I2.14744.

[28] A. A. Farisi, Y. Sibaroni, and S. Al Faraby, "Sentiment analysis on hotel reviews using multinomial naïve Bayes classifier," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, May 2019, pp. 1–10. doi: 10.1088/1742-6596/1192/1/012024.

[29] N. S. M. Nafis and S. Awang, "An enhanced hybrid feature selection technique using term frequency-inverse document frequency and support vector machine-recursive feature elimination for sentiment classification," *IEEE Access*, vol. 9, pp. 52177–52192, 2021, doi: 10.1109/ACCESS.2021.3069001.

[30] S. W. A. Sherazi, J. W. Bae, and J. Y. Lee, "A soft voting ensemble classifier for early prediction and diagnosis of occurrences of major adverse cardiovascular events for STEMI and NSTEMI during 2-year follow-up in patients with acute coronary syndrome," *PLoS One*, vol. 16, no. 6, Jun. 2021, doi: 10.1371/journal.pone.0249338.

[31] D. Musfiroh, U. Khaira, P. E. P. Utomo, and T. Suratno, "Analisis sentimen terhadap perkuliahan daring di Indonesia dari Twitter dataset menggunakan InSet lexicon," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 1, no. 1, pp. 24–33, 2021.

[32] J. P. D. Delizo, M. B. Abisado, and Ma. I. P. D. L. Trinos, "Philippine Twitter sentiments during Covid-19 pandemic using multinomial naïve Bayes," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 1.3, pp. 408–412, Jun. 2020, doi: 10.30534/ijatcse/2020/6491.32020.

[33] F. S. Khurniawan and Y. Ruldeviyani, "Twitter sentiment analysis case study on the revision of the Indonesia's corruption eradication commision (KPK) law 2019," in *2020 International Conference on Data Science and Its Applications (ICoDSA)*, 2020, pp. 1–6. doi: 10.1109/ICoDSA50139.2020.9212851.

[34] T. Sontayasara *et al.*, "Twitter sentiment analysis of bangkok tourism during Covid-19 pandemic using support vector machine algorithm," *Journal of Disaster Research*, vol. 16, no. 1, pp. 24–30, 2021, doi: 10.20965/jdr.2021.p0024.

[35] A. T. Mahmood, S. S. Kamaruddin, R. K. Naser, and M. M. Nadzir, "A combination of lexicon and machine learning approaches for sentiment analysis on facebook," *Journal of System and Management Sciences*, vol. 10, no. 3, pp. 140–150, 2020, doi: 10.33168/JSMS.2020.0310.

[36] S. Pradha, M. N. Halgamuge, and N. T. Q. Vinh, "Effective text data preprocessing technique for sentiment analysis in social media data," in *11th International Conference on Knowledge and Systems Engineering*, 2019, pp. 1–8. doi: 10.1109/KSE.2019.8919368.