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

Salsabila Rahadatul Aisy¹, Budi Prasetiyo²
¹²Computer 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 naïve 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 naïve 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

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.
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.
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 "naah" 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 ith 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

Figure 1. Flowchart Research Design
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-dimension 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 hyperplane 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>
<thead>
<tr>
<th>Actual</th>
<th>Prediction</th>
<th>Positive</th>
<th>Negative</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>TP (True Positive)</td>
<td>FN (False Negative)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>FP (False Positive)</td>
<td>TN (True Negative)</td>
<td></td>
</tr>
</tbody>
</table>

\[
\text{accuracy} = \frac{TP+TN}{TP+FP+FN} \times 100\% \quad (1)
\]

**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
Recursive Journal of Informatics, Vol. 1, No. 2, September 2023 | 97

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.

![Figure 2. Extend stop words list](image)

**Table 2. Result of preprocessing**

<table>
<thead>
<tr>
<th>Tweet</th>
<th>Data cleaning</th>
<th>Tokenization</th>
<th>Normalization</th>
<th>Stop words removal</th>
<th>Stemming</th>
</tr>
</thead>
<tbody>
<tr>
<td>Puan Maharani Akan Pimpin Rapat Paripurna RUU TPKS Menjadi UU <a href="https://t.co/gLz7HM4i">https://t.co/gLz7HM4i</a> NC</td>
<td>puhan maharani akan pimpin rapat paripurna ruu tpsks menjadi uu</td>
<td>[puhan, maharani, akan, pimpin, rapat, paripurna, ruu, tpsks, menjadi, uu]</td>
<td>[puhan, maharani, akan, pimpin, rapat, paripurna, ruu, tpsks, menjadi, uu]</td>
<td>[pimpin, paripurna]</td>
<td>[pimpin, paripurna]</td>
</tr>
<tr>
<td>@andianisharfina @maida__Udah bisa pake nama #UUTPKS udah sah</td>
<td>udah bisa pake nama udah sah</td>
<td>[udah, bisa, pake, nama, udah, sah]</td>
<td>[sudah, bisa, pakai, nama-nama, sudah, sah]</td>
<td>[pakai, nama-nama, sah]</td>
<td>[pakai, nama, sah]</td>
</tr>
<tr>
<td>... @rizkiwulandani @byeoreum @tegarftr @dikignwn @Wirogendheng88 @tubrffes UU TPKS dan kenaikan usia menikah aja udah ngebuktiin ini salah child marriage itu dihalalkan di islam sedangkan batas usia perkahwinan itu disini</td>
<td>uu tpsks dan kenaikan usia menikah aja udah ngebuktiin ini salah child marriage itu dihalalkan di islam sedangkan batas usia perkahwinan itu disini</td>
<td>[uu, tpsks, dan, kenaikan, usia, menikah, aja, udah, ngebuktiin, ini, salah, child, marriage, itu, dihalalkan, di, islam, sedangkan, batas, usia, perkahwinan, itu, disini]</td>
<td>[uu, tpsks, dan, kenaikan, usia, menikah, aja, udah, ngebuktiin, ini, salah, child, marriage, itu, dihalalkan, di, islam, sedangkan, batas, usia, perkahwinan, itu, disini]</td>
<td>[kenaikan, usia, menikah, membuktikan, salah, child, marriage, dihalalkan, islam, batas, usia, perkahwinan]</td>
<td>[naik, usia, meni, bukti, salah, child, marriage, halal, islam, batas, usia, nikah]</td>
</tr>
<tr>
<td>UU TPKS Disahkan, Berikut 6 Poin Penting dari Permenkabdri <a href="https://t.co/rtxsuhD4o">https://t.co/rtxsuhD4o</a> x</td>
<td>uu tpsks disahkan berikut poin penting dari permenkabdri</td>
<td>[uu, tpsks, disahkan, berikut, poin, penting, dari, permenkabdri]</td>
<td>[uu, tpsks, disahkan, berikut, poin, penting, dari, permenkabdri]</td>
<td>[disahkan, poin, permendikbudristek]</td>
<td>[sah, poin, permendikbudristek]</td>
</tr>
</tbody>
</table>

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.
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.

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.
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

<table>
<thead>
<tr>
<th>Authors</th>
<th>Data</th>
<th>Language</th>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delizo et al. (2020)</td>
<td>Twitter</td>
<td>Tagalog</td>
<td>Multinomial naïve Bayes</td>
<td>72.00%</td>
</tr>
<tr>
<td>Asif et al. (2020)</td>
<td>Facebook</td>
<td>Multilingual</td>
<td>Multinomial naïve Bayes</td>
<td>66.00%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Support vector machine</td>
<td>82.10%</td>
</tr>
<tr>
<td>Khurmanaw &amp; Ruđeviyan, (2020)</td>
<td>Twitter</td>
<td>Indonesia</td>
<td>Support vector machine</td>
<td>81.70%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Naïve Bayes</td>
<td>80.90%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Decision tree</td>
<td>74.55%</td>
</tr>
<tr>
<td>Sontayasara et al. (2021)</td>
<td>Twitter</td>
<td>English</td>
<td>Support vector machine</td>
<td>71.00%</td>
</tr>
<tr>
<td>Andrián et al. (2022)</td>
<td>Twitter</td>
<td>Indonesia</td>
<td>Hard voting</td>
<td>75.43%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Soft voting</td>
<td>75.86%</td>
</tr>
<tr>
<td>Proposed method</td>
<td>Twitter</td>
<td>Indonesia</td>
<td>Soft voting</td>
<td>84.31%</td>
</tr>
</tbody>
</table>

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


