**Sentiment Analysis Regarding the TPKS Law on Twitter Social Media Using the Support Vector Machine and K-Nearest Neighbor Methods**

Arviandri Naufal Zaki, Dian Pratiwi, ST, MTI, and Syandra Sari, S.Kom, M.Kom

123Department of Informatics Engineering, Faculty of Industrial Technology, Trisakti University, Jl. Letjen S. Parman No.1, RT.6/RW.16 Grogol, Kec. Grogol Petamburan, West Jakarta City, Special Capital Region of Jakarta, 11440

*E-mail:* [*arviandri064001800035@std.trisakti.ac.id*](mailto:arviandri064001800035@std.trisakti.ac.id)

**Abstract**

Twitter is a social media that is widely used by the people of Indonesia and the world. Twitter is also used to share news and personal opinions, market products, to criticize a policy or regulation. Opinions posted as tweets on Twitter can also be used as a benchmark for whether the policies issued are widely supported or otherwise. To obtain these benchmarks, sentiment analysis is used to separate positive opinions from negative opinions. For data retrieval, scraping from the Twitter website is used to get it. After that, the initial process is carried out before the data is processed, namely Preprocessing to eliminate parts that are not useful in data. Then the Support Vector Machine and K-Nearest Neighbor techniques were carried out to classify positive and negative opinions in order to compare which one had more opinions and then be used as a benchmark for a policy.

**Keywords:** *Sentiment Analysis, UU TPKS, Support Vector Machine, K-Nearest Neighbour, Scraping*

1. **Introduction**

Lately, many crimes of sexual violence occurred in society, according to Komnas Perempuan, it was recorded that there was 14,719 cases of sexual violence against women throughout year 2020. Actually, the government has drafted a law on criminal acts of sexual violence since 2016, but it was only passed in 2022 after pressure from various parties, the ratification of the TPKS Law raises many pros and cons and causes people to have opinions in that on social media. One of those who support argues that this law will make perpetrators of sexual crimes can be punished according to what they did, while one of those who oppose argues that this law will create an idea that free sex is allowed. One of the social media that is often used to express opinions by the Indonesian people is Twitter.

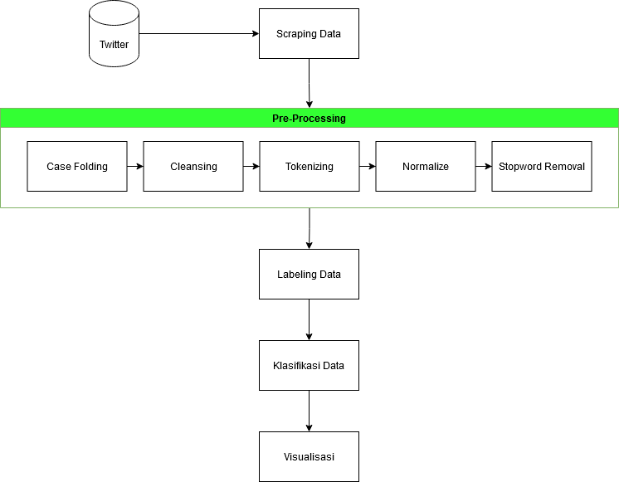
Twitter is used by the Indonesian people for various things, such as communicating with other people publicly or privately, sharing news and personal opinions, selling, to criticizing or praising something. Because the information on Twitter is also limited to around 280 characters, users usually only send something short [1].

The government also uses this platform to find out the public's response to the newly issued policies as in this study. Therefore, Twitter users can have opinions about the policies issued are influenced by emotions that can be classified to determine the polarization, namely positive or negative about tweets about government policies in this study.

Sentiment analysis is the activity of processing words to produce a sentiment (positive or negative). Sentiment analysis aims one of which is to get an opinion from a newly issued government policy then the opinion is classified into positive and negative sentiments. The technique used to retrieve data from Twitter before being analyzed is using the Scraping technique, which is taking data directly from the Twitter website. Then the technique used to provide sentiment (label) is Valence Aware Dictionary and Sentiment Reasoner (VADER) because this technique has high accuracy and can be used for negation, then the technique used to classify the data is K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) because this two classification models have their respective advantages in previous studies regarding the comparison of the accuracy of the two, the accuracy level is quite high for SVM (89.7%) while KNN is able to process large data in a short time (1,113 data within 0.0160 seconds) [2].

1. **Method**

When compiling the final project, the author uses a flowchart as follows:



**Fig 1.** Flowchart from this Research

The steps needed to complete this research are as follows :

* 1. **Data Retrieval**

Retrieval data from one of the largest social media, i.e. Twitter, uses data scraping techniques from a library. The data collected is in the form of Indonesian-language tweets with the keyword "uu tpks" in the period of 12 April 2022 to 24 April 2022 and retweets are not included.

* 1. **Data Pre-Processing**

After collecting data before being analyzed, it is necessary to carry out an initial process or known as Preprocessing. This process will process the initial data that is still irregular to be used as regular data that can be applied to the next process. The preprocessing consists of Case Folding, Cleansing, Tokenizing, Normalization, and Stopword Removing.

* + 1. **Case Folding**

Case Folding is a step to change the uppercase letters or capital letters (uppercase) contained in the text into lowercase letters (lowercase).

* + 1. **Case Folding**

Cleansing is a step to clean data from unnecessary things such as URLs, hashtags, punctuation marks, numbers and so on.

* + 1. **Tokenizing**

Tokenizing is making changes from a word in a sentence separated by a separator (space) into a token.

* + 1. **Normalize**

Normalization is a process in which non-standard or short words are converted into correct standard words.

* + 1. **Stopword Removing**

Stopword Removing is a process where conjunctions such as which, in, to, from which are not needed in the analysis process are removed.

* 1. **Data Labeling**

After the data is cleaned then labeling is done on the data. Labeling on the data is done automatically using a dictionary that already contains the sentiment weight (lexicon) and the total is calculated from the sentiment based on the total weight of all words in each data. In this research, for this research, labeling data using the Vader method.

Lexicon Based Features is a word feature that contains positive and negative sentiments based on a dictionary (lexicon). Lexicon is a collection of words on known sentiments and collected in the form of a dataset [3].

Vader is a method or tool in conducting sentiment analysis based on lexicon or rules that have been made to approach sentiment on social media [4].

* 1. **Word weighting**

After being given a label, the word weighting is carried out. Word weighting is done using Term Frequency (TF) and Inverse Document Frequency (IDF).

TF-IDF is an algorithm that can generate information about how often the word appears in the dataset and is displayed in the form of weight per word. To determine the weight per word, this algorithm uses several components that match the name, namely Term Frequency (TF) and Inverse Document Frequency (IDF) [5].

* 1. **Classify the Data**

This process aims to process data into positive opinions and negative opinions. There are many methods to classify data, one of which is Support Vector Machine and K-Nearest Neighbor. Is one method for classifying data and regression. In this study, the author uses the Support Vector Machine and K-Nearest Neighbor methods to classify the data.

Support Vector Machine is a classification method that uses a linear classification method by finding the best hyperlane that functions as a separator between 2 classes. The basic principle is that linear classification is carried out and then developed so that it can be used in non-linear problems by incorporating the concept of kernel trick in high-dimensional workspaces [6].

K-Nearest Neighbor is an algorithm for classification that uses a method of measuring the level of similarity between neighboring data (cosine similarity) or measuring the Euclidean distance from training data (training data) with test data (test data) [7].

* 1. **Visualization**

In this process, visualization of the data generated from the classification process will be carried out. The purpose of this process is to make it easier to read the intent and information from the results of the analysis.

* 1. **Checking the Accuration using K-Fold Cross Validation**

K-Fold Cross Validation is a method to sort the initial data into test data and train data. This method is done so that there is no bias in sampling (test data). This method works by dividing train data with test data continuously so that each data has the opportunity to become test data [8].

1. **Implementation and Result**
   1. **Data Retrieval / Data Scraping**

At this stage the tweet data on Twitter is taken using the Scrapping Technique. The library used at this stage is the *snscrape* library. *Snscrape* is a library that contains several functions that can be used to pull data from social media such as Facebook, Instagram, Twitter, and so on.

This process resulted in 15,632 tweet data ranging from 12 to 24 April 2022. The date was taken because that was the date when the TPKS Bill was passed into the TPKS Law.

* 1. **Pre - Processing**

At this stage, the cleaning and modification of the data that has been previously collected is carried out so that there is no data that can interfere with the analysis and to make the data can be processed to the next stage.

At the pre-processing stage there are 5 stages used to make the data clean and ready to be used for the next stage, these stages are:

* + 1. Case Folding

This stage aims to change capital letters to lowercase letters so that the data is evenly distributed.

Examples of case folding in this research are as follows :

**Table 1.** An example of the case folding process

|  |  |
| --- | --- |
| Before ”*Case Folding*” | After ”*Case Folding*” |
| Lihat tanggal chatnya, kalau setelah April udah bisa dijerat UU TPKS | lihat tanggal chatnya, kalau setelah april udah bisa dijerat uu tpks |

* + 1. Cleansing

This stage aims to eliminate hashtag data, mentions, punctuation marks, numbers, urls, useless spaces, and non-ASCII data such as emoticons, Chinese data, and so on.

Examples of cleansing stage in this research are as follows :

**Table 2.** An example of the cleansing process

|  |  |
| --- | --- |
| Before ”*Cleansing*” | After ”*Cleansing*” |
| lihat tanggal chatnya, kalau setelah april udah bisa dijerat uu tpks 😅 https://t.co/c89NcYzjQv | lihat tanggal chatnya, kalau setelah april udah bisa dijerat uu tpks |
| UUTPKS diterapkan keras mulai dari pemrentah dan @DPR\_RI, setuju? @KemensetnegRI https://t.co/isAu9nBZpx | UUTPKS diterapkan keras mulai dari pemrentah dan setuju |

* + 1. Tokenizing

At this stage the data will be separated by separator (space) into tokens (words in each sentence in the data). This step serves to make the data compatible with the next steps such as normalize, stopword removal, and so on.

**Table 3.** An example of the tokenizing process

|  |  |
| --- | --- |
| Before”*Tokenizing*” | After ”*Tokenizing*” |
| lihat tanggal chatnya kalau setelah april udah bisa dijerat uu tpks | [lihat, tanggal, chatnya, kalau, setelah, april, udah, bisa, dijerat, uu, tpks] |

* + 1. Normalize

At this stage, non-standard data are converted into standard words. This step is taken to prevent the occurrence of words that are outside the lexicon (out of vocabulary) because most of the lexicon are standard words.

**Table 4.** An example of the normalize process

|  |  |
| --- | --- |
| Before ”*Normalize*” | After ”*Normalize*” |
| kagak ada yg ribut nyuruh seragaman baju nasional kebaya  uu tpks sah  vaksin serviks bakal jd vaksin wajib | tidak ada yang ribut nyuruh seragaman baju nasional kebaya uu tpks sah vaksin serviks bakal jadi vaksin wajib |

* + 1. Stopword Removal

This stage aims to eliminate unnecessary words in the data that can interfere with the analysis. The stopword list file used is obtained from the spaCy library data which is combined with a other stopword list on github.

**Table 5.** An example of the stopword removal process

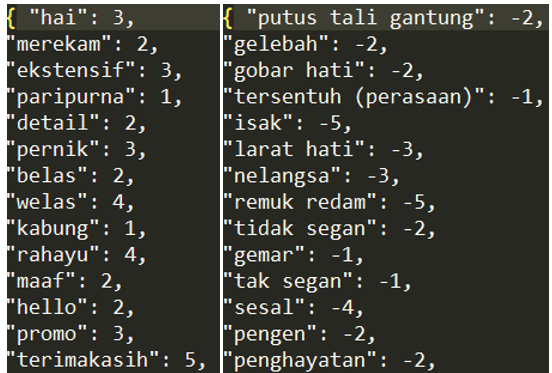
|  |  |
| --- | --- |
| Sebelum ”*Stopword Removal*” | Setelah ”*Stopword Removal*” |
| lihat tanggal chatnya kalau setelah april sudah bisa dijerat uu tpks | tanggal chatnya april dijerat uu tpks |

* 1. **Data Labelling**

After the data is cleaned through the pre-processing stage, then it enters the data labeling stage. This stage aims to label the data into a sentiment (positive & negative). The labeling used is a Lexicon-based method with the VADER (Valence Aware Dictionary and Sentiment Reasoner) library.

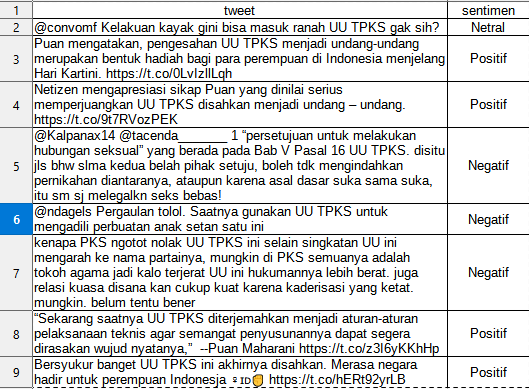
The steps contained in this stage are the data from the previous stage (Pre-Processing) which is still in the form of a token, first it is un-tokenized so that it can be processed by the library, after that the default dictionary from the VADER library is cleaned first and replaced by Inset dictionary and then labeled for each row based on the polarity value generated by the VADER library after the labeling is complete and then visualized using Wordcloud and Pie Chart.

An example content of the Inset (Indonesian Sentiment) Lexicon is as follows :

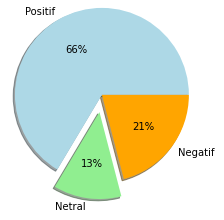


**Fig 2.** An example content of Inset (left for positive lexicon and right for negatif lexicon)

And for the result of the this process is as follows :



**Fig 3.** An example result of the data labeling process



**Fig 4.** Pie chart from the result of data labeling



**Fig 5.** Wordcloud from the data labeling process (Positive Data)



**Fig 6.** Wordcloud from the data labeling process (Negative Data)

It can be seen from Figure 4 that the public's sentiment regarding the TPKS Law is mostly positive at 66% (10,385 data). And from Figure 5 & 6 it can be concluded that the words that appear the most are "Kekerasan" and "Seksual" and there are also collocation words such as "Perempuan indonesia" and "Terima kasih".

* 1. **Data Weighting**

At this stage, a process is to convert the collected words into vectors so that the data can be used in the next process, which is the classification process using SVM and KNN.

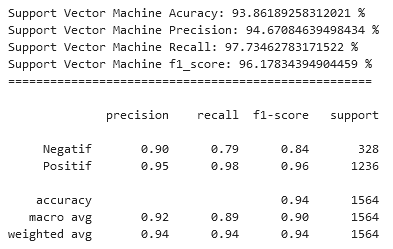
At this stage this research uses the TF-IDF method.

* 1. **Classify the Data using SVM Method**

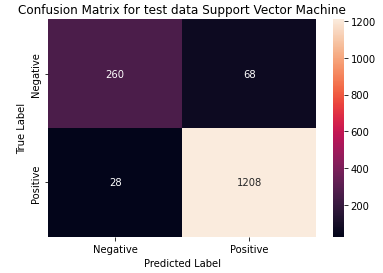
At this stage the data that has gone through the data weighting process, the next data enters the classification stage. This stage aims to classify the data based on the sentiments that have been obtained previously.

The classification is carried out using a linear kernel because it has widely used in many research and in previous research the method was quite accurate.

After classifying the train data, a visualization is displayed in the form of the accuracy of the classification carried out in the previous process. The visualization of the data generated by this method is as follows :



**Fig 7.** Results of Linear SVM Accuracy Visualization

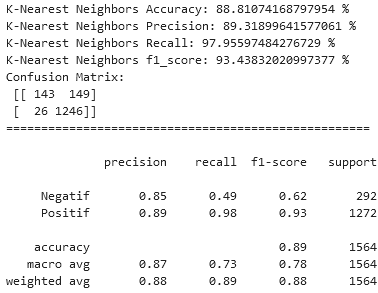


**Fig 8.** Results of Visualization Using Confusion Matrix in SVM Linear

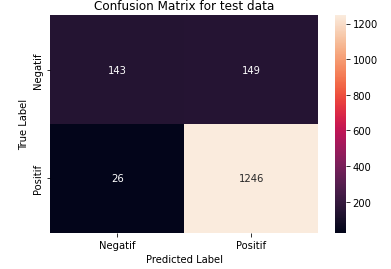
* 1. **Classify the Data using KNN Method**

At this stage, the same as the previous stage of data that has gone through the data weighting process, the data then enters the classification stage. This stage aims to classify the data based on the sentiments that have been obtained previously. At this stage the classification is carried out using the K-Nearest Neighbor (KNN) method.

After modeling the train data and classifying the test data, a visualization is displayed in the form of the accuracy of the classification carried out in the previous process. The visualization of the data generated by this method is as follows :



**Fig 9.** Results of KNN Accuracy Visualization



**Fig 10.** Results of Visualization Using Confusion Matrix inSVM Linear

* 1. **Accuration Checking using K-Fold Cross Validation**

At this stage the previously generated accuracy is checked using the K-Fold Cross Validation method due to reducing bias in the distribution of train data and test data.

From the process of checking the accuracy using K-Fold Cross Validation, an accuracy of 93.2% was obtained for SVM Linear and 87.2% for KNN.

1. **Conclusion**

The data obtained in the data scrapping is 15,632 data. The results of the sentiment obtained using Vader in this study are positive by 66% (10,385 data), negative by 21% (3,274 data) and neutral by 13% (1,973 data). From these results, it can be concluded that it shows public satisfaction with the policies issued.

In the classification process, it was found that the best iteration KNN classification for this study was 4 times to produce a fairly accurate accuracy. And the conclusion of the classification is that the SVM method is more accurate than the KNN method with an accuracy of 92% for SVM and 88% for KNN. And after being checked using K-Fold cross validation the accuracy produced is the same, meaning that there are no biased settings in this research.

**References**

[1] K. Makice, *Twitter API: Up and Running*. 2009.

[2] M. R. A. Nasution and M. Hayaty, “Perbandingan Akurasi dan Waktu Proses Algoritma K-NN dan SVM dalam Analisis Sentimen Twitter,” *J. Inform.*, vol. 6, no. 2, pp. 226–235, 2019, doi: 10.31311/ji.v6i2.5129.

[3] M. Desai and M. A. Mehta, “Techniques for sentiment analysis of Twitter data: A comprehensive survey,” *Proceeding - IEEE Int. Conf. Comput. Commun. Autom. ICCCA 2016*, pp. 149–154, 2017, doi: 10.1109/CCAA.2016.7813707.

[4] S. Elbagir and J. Yang, “Language Toolkit and VADER Sentiment,” *Proc. Int. MultiConference Eng. Comput. Sci.*, vol. 0958, pp. 12–16, 2019.

[5] J. Patterson and A. Gibson, *Deep learning: A Practionar Approach*, vol. 521, no. 7553. 2017. [Online]. Available: http://www.nature.com/doifinder/10.1038/nature14539

[6] R. Feldman and J. Sanger, *The Text Mining Handbook*. 2007.

[7] F. Gorunescu, *Data Mining Concepts,Models and Techniques*. 2011.

[8] T. Ridwansyah, “Implementasi Text Mining Terhadap Analisis Sentimen Masyarakat Dunia Di Twitter Terhadap Kota Medan Menggunakan K-Fold Cross Validation Dan Naïve Bayes Classifier,” *KLIK Kaji. Ilm. Inform. dan Komput.*, vol. 2, no. 5, pp. 178–185, 2022, [Online]. Available: https://djournals.com/klik