Analisis Sentimen Kemendikbud menggunakan Vader dan RBF, Polynomial, Linier Kernel SVM Berbasis Binary Particle Swarm Optimization

Sentiment Analysis of the Minister of Education and Culture using Vader and RBF, Polynomial, Linier Kernels SVM Based on Binary Particle Swarm Optimization

[1]Rutlima Sinaga, [2]Ilham Firman Ashari*, [3]Winda Yulita
[1,2,3]Program Studi Teknik Informatika, Jurusan Teknologi Produksi dan Industri, Institut Teknologi Sumatera
Jl. Terusan Ryacudu. Lampung Selatan
*e-mail: firman.ashari@if.itera.ac.id

(***received*:** 19 Juli 2022, ***revised*:** 27 April 2023, ***accepted*:** 28 April 2024)

Abstrak

Komentar dari media sosial dapat dianalisis lebih lanjut. Media sosial digunakan untuk berinteraksi antara satu orang dengan orang lain, maupun dengan pemerintah. Isu ini mengemuka karena adanya perdebatan dan opini publik dari masyarakat, lembaga, dan LSM terkait Peraturan Menteri Nomor 30 Tahun 2021 tentang Pencegahan dan Penanganan Kekerasan Seksual. Di Lingkungan Pendidikan Tinggi Oleh karena itu Dalam Penelitian Ini Kami Ingin Mengkaji Apa Yang Menjadi Akar Utama Permasalahan Dengan Menggunakan Pendekatan Metodis Dengan Menggunakan Natural Language Processing. Pra-pemrosesan yang diterapkan adalah pelipatan kasus, tokenisasi, penghapusan stop word, stemming menggunakan literatur. Model yang mengimplementasikan PSO gagal meningkatkan akurasi pada semua kernel. Performa terbaik sebelum menerapkan PSO ke dataset Twitter menggunakan kernel linier. Penelitian ini melakukan analisis sentimen terhadap terbitnya Peraturan Menteri No. 30 Tahun 2021. Data yang diperoleh kemudian diolah terlebih dahulu. Performa yang diukur adalah akurasi dan f1-makro pada model tanpa PSO dan akurasi pada model menggunakan akurasi. Model yang akan dibentuk menggunakan kernel linier, RBF dan polinomial orde 1 dan orde 2. Analisis kalimat merupakan bidang yang menganalisis sentimen, sikap dan emosi entitas serta atributnya dalam bentuk teks. Tujuan dari penelitian ini adalah untuk membandingkan kinerja algoritma klasifikasi Support Vector Machine tanpa pemilihan fitur Particle Swarm Optimization dan kinerja algoritma klasifikasi Support Vector Machine menggunakan pemilihan fitur Particle Swarm Optimization. Data yang diperoleh kemudian diolah terlebih dahulu. Kumpulan data secara otomatis diberi label menggunakan VADER (Valence Dictionary for Sentiment Reasoning). Kernel yang berhasil meningkatkan akurasi adalah kernel RBF dan polinomial pada dataset Twitter.

Kata Kunci: SVM, Vader, PSO, Analisis Sentimen, Kebijakan Pemerintah

Abstract

Comments from social media can be analyzed further. Social media is used to interact from one person to another, as well as with the government. This Issue Was Raised Because Of Debate And Public Opinion From The Community, Institutions And Ngos Regarding Ministerial Regulation No. 30 Of 2021 Concerning Prevention And Handling Of Sexual Violence. In The Higher Education Environment, Therefore In This Research We Want To Examine What Is The Main Root Of The Problem Using A Methodical Approach Using Natural Language Processing. The pre-processing applied is case folding, tokenization, elimination of stop words, stemming using literature. The model implementing PSO failed to improve accuracy on all kernels. Best performance before applying PSO to twitter dataset using linear kernel. This study conducted sentiment analysis regarding the issuance of ministerial regulation no. 30 of 2021. The data obtained was then preprocessed. The performance measured is accuracy and f1-macro in the model without PSO and accuracy in the model using

 ISSN:2302-8149
e-ISSN:2540-9719

accuracy. The model to be formed uses linear kernels, RBF and polynomials of order 1 and order 2. Sentence analysis is a field that analyzes sentiment, attitudes and emotions of entities and their attributes in text form. The aim of this research is to compare the performance of the Support Vector Machine classification algorithm without Particle Swarm Optimization feature selection and the performance of the Support Vector Machine classification algorithm using Particle Swarm Optimization feature selection. The data obtained is then pre-processed. The data set was automatically labeled using VADER (Valence Dictionary for Sentiment Reasoning). The kernels that succeeded in increasing accuracy were the RBF kernel and polynomials on the Twitter dataset.

Keywords: *SVM, Vader, PSO, Sentiment Analysis, Government Policy*

1 Introduction

The development of the internet has introduced online media, such as social media, which is used massively. Social media is a means of interaction between users by creating, sharing and receiving information content through internet-based applications. The government also uses social media to interact and convey relevant information to the public. Based on data collected from the Indonesian Internet Network Penetration Association (APJII), internet users in Indonesia in 2019-2021 reached 77.3% of the total percentage of Indonesian society, equivalent to around 198,714,074 million people. This is in line with the contents of Law Number 14 of 2008 concerning Openness of Public Information which emphasizes the importance of access to public information for the community. With the significant growth of internet users, social media has become an important platform in facilitating interaction and dissemination of information related to the public interest [1].

The issue that is being widely discussed is the regulation that has just been announced by the Ministry of Education, Culture, Research, Technology and Higher Education, namely Permendikbud Ristekdikti Number 30 of 2021 concerning the Prevention and Handling of Sexual Violence in Higher Education Environments. Based on the annual report of the National Commission on Violence Against Women, it is noted that cases of sexual violence tend to increase from year to year [2]. However, currently existing law enforcement has not been able to provide protection and justice for victims of sexual violence. The sanctions given to perpetrators also do not reflect appropriate law enforcement [3]. One institution that reports many cases of sexual violence is universities.

This regulation was officially announced on August 31 2021 and then promulgated on September 3 2021 in Jakarta. His focus on students fuelled the research conducted at ITERA. Not only that, the response from the public and netizens on social media was no less enthusiastic in commenting on the substance of the regulation. Articles 3 and 5, which touch on the prevention and handling of sexual violence, are the main focus in discussions surrounding the contents of the regulations. This issue even managed to occupy the top position in trending Twitter topics on November 4 2021. With so many opinions appearing, it is difficult for anyone to read each one. Therefore, a sentiment analysis was carried out to see a picture of public opinion regarding the issuance of Permendikbud Number 30 of 2021. Through this sentiment analysis, the sentiments, attitudes and emotions of each text that were revealed were examined [4][5].

In classifying text based on sentiment, the first step is to weight the text data into numerical form, allowing the subsequent use of machine learning techniques. This process is known as feature extraction [6]. There are a number of techniques available, including unigrams, bigrams, trigrams, term frequency-inverse document frequency (TF-IDF), and Word Embedding. In previous research, analysis concluded that TF-IDF was a more effective method. TF-IDF is proven to be superior to the one-hot encoding method, word2vec, and paragraph2vec [7] [8]. In fact, the research results confirm that TF-IDF is also superior to the bag of words and n-gram approaches. The importance of feature extraction in the sentiment classification process is rooted in the conversion of text into a numerical representation, allowing machine learning algorithms to process it more efficiently. The use of techniques such as unigrams, bigrams, and trigrams allows for a deeper understanding of the context of words. However, TF-IDF, which combines the frequency of occurrence of a word in a document with its presence in the entire corpus, was shown to outperform these alternatives.

TF-IDF, as a feature extraction method, offers a careful evaluation of the words in the text, giving them appropriate weight according to their uniqueness [9]. Its accuracy in assessing and distinguishing important words from less relevant ones brings significant benefits in sentiment

 ISSN:2302-8149
e-ISSN:2540-9719

analysis. In this context, TF-IDF is proven to be more capable of differentiating and interpreting sentiments more accurately than other approaches that assume each word has similar weights. The accuracy and thoroughness of TF-IDF in assessing the meaning of words is the main reason why this method is a careful choice in feature extraction for sentiment analysis. Its consistent ability to process text and capture the essence of the context of words provides a strong foundation for its use in sentiment research. TF-IDF not only simplifies the sentiment classification process, but also strengthens its analysis by placing emphasis on words that have substantial impact. Consistency and accuracy in sorting words and giving appropriate weight to their contribution in the context of the text makes TF-IDF a wise choice in processing text data for deeper and more accurate sentiment analysis.

This research, based on a literature review and consideration of the reliability of machine learning algorithms, focuses on sentiment analysis regarding PERMENDIKBUD No. 30 of 2021. In this effort, the Support Vector Machine algorithm based on Particle Swarm Optimization is used for feature extraction and selection. The aim is to evaluate sentiment towards regulations issued by the Ministry of Education, Culture, Research, Technology and Higher Education.

2 Literature Review

Research in Sentiment Analysis has achieved many achievements, with a variety of different methods and datasets. For example, in the Sentiment Analysis research regarding Lockdown on Twitter in 2020, an analysis was carried out of people's reactions to the lockdown policy. Using VADER for data labeling and TF-IDF for weighting followed by classification using Naïve Bayes and Support Vector Machine, achieved accuracies of 81% and 87% respectively [10]. Tati Mardiana conducted an analysis regarding Franchise Businesses on Twitter, comparing the Neural Network, K-Nearest Neighbor, Support Vector Machine, and Decision Tree methods. The results show the superiority of SVM with an accuracy of 83% and an AUC value of 0.879 [11]. Satria Yudha and his colleagues evaluated classification algorithms for Indonesian language film reviews, finding that SVM achieved an AUC value of 0.986 and an accuracy of 93.57% on German language reviews for the film Avengers: Infinity War [12][13]. Research by Pande Made R C D and his team aims to compare term weighting and word embedding techniques in local government short text classification. The results show the superiority of the combination of TF-IDF with SVM linear kernel compared to Logistic Regression [14].

There are many types of algorithms for classifying sentiment, one of which is Support Vector Machine (SVM) [15]. SVM in several studies achieved higher accuracy [12][16]. However, handling text data which is unstructured data requires many attributes during the classification process. These factors make the classification process heavier or affect accuracy [17]. The solution to this problem is to select only important features, namely feature selection. Feature selection is a stage in data processing that can influence the level of accuracy or improve the base classifier. Four examples of Feature Selection include: (1) genetic algorithm, (2) evolutionary programming, (3) evolutionary strategies and genetic programming, (4) PSO [18][15][19]. The PSO Selection feature is easy to implement and can find optimal points quickly [20]. Research conducted by Siti Ernawati, et al compared feature selection Genetic Algorithm (GA) with PSO using the Naïve Bayess method. The combination of PSO and Naïve Bayes is better with an accuracy of 98.00% [21]. Previous research compared GA and PSO, both of which were successfully implemented. PSO is better than GA in reducing the number of features [22].

3 Research Methods

The stages in the process flow include data collection, preprocessing, feature extraction, split data, classification, evaluation of classification models, and analysis of research results. The research stages can be seen in Figure 1.

3.1 Data Collection

Data collection is the first step in obtaining the required data [23]. Data is a crucial element in sentiment analysis. In this research, data was first obtained from the social media platform Twitter using the scrapping technique. However, to overcome data limitations related to the issue that is the focus of the research, additional data sources are needed. In addition, the model used initially had a

 ISSN:2302-8149
e-ISSN:2540-9719

low level of accuracy, so it needed to be improved. To increase accuracy, researchers decided to expand the data used in the learning process. The following are the steps taken in retrieving data from the Twitter.

3.2 Scrape Twitter

Twitter scraping was carried out in a Jupyter notebook using the tweepy library for data retrieval. The tweepy library utilizes the API provided by Twitter. The following are the steps required in the process:

1. Registration for a personal Twitter account is done via the official developer.twitter.com page. After the developer account is approved, tokens such as consumer key, consumer secret, access token, and access token secret will be obtained.
2. Use the token to authenticate your Twitter account for verification as a developer.
3. Determining the objects to be taken in the scraping process. Tweepy provides various objects such as tweet.user, tweet.full.text, tweet.text, tweet.created_at, tweet.id_str, and others that can be loaded from the Twitter platform.

In this study, tweet.full.text, tweet.lang, tw eet.user.screen_name were used.

The data preprocessing flow is depicted in Figure 1 below and the preprocessing stages in the form of examples can be seen in the table 1.

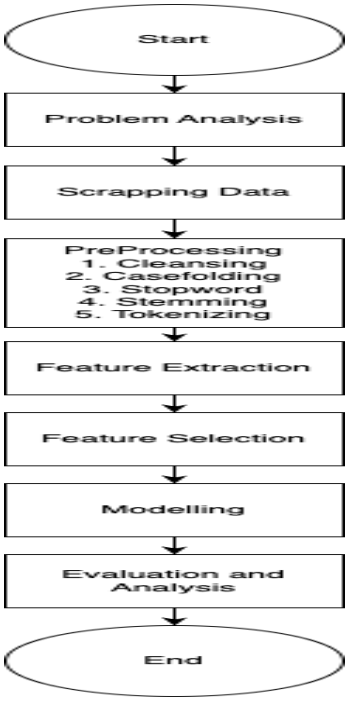

Figure 1. Research Workflow

Table 1. Dataset Preprocessing Stage On Indonesian Tweets

Preprocessing	Before	After
Cleansing -> removal of noise such as punctuation, urls, emoticons	Kita lanjutkan saja diam ini, hingga dirimu dan diriku mengerti. tidak semua kebersamaan, harus melibatkan hati.	Kita lanjutkan saja diam ini hingga dirimu dan diriku mengerti tidak semua kebersamaan harus melibatkan hati
Case folding-> changes the text to lower case	Kita lanjutkan saja diam ini hingga dirimu dan diriku mengerti tidak semua kebersamaan harus melibatkan hati	kita lanjutkan saja diam ini hingga dirimu dan diriku mengerti tidak semua kebersamaan harus melibatkan hati
Stopword-> reduces common words that often appear and are	kita lanjutkan saja diam ini hingga dirimu dan diriku mengerti tidak	lanjutkan diam mengerti tidak semua kebersamaan harus

 ISSN:2302-8149
e-ISSN:2540-9719

not relevant and deletes words that consist of less than 4 letters such as the words "in", "and", "I", "this", "that"	semua kebersamaan harus melibatkan hati	melibatkan hati
Tokenize -> divides text into tokens	lanjut saja diam ini erti tidak semua sama harus libat hati	“lanjut” “diam” “erti” “tidak” “semua” “sama” “harus” “libat” “hati”

Data that has gone through all stages will be used as terms or tokens. In table 2 it is found that the tokens in each row of data are not the same, this is common because of the diversity of opinions from users.

Table 2. Pre-Processing Result

Document	Result
D1	“lanjut” “diam” “erti” “tidak” “semua” “sama” “harus” “libat” “hati”
D2	“lagu” “bosan”
D3	“makasih” “hasil”
D4	“nomor” “satu” “bukan”

The collection of terms is then applied by the join method to combine each word into one sentence on each line. The results of this process will be used in the next stage, namely labeling the dataset using the Vader library.

3.3 Dataset Labelling

The labeling stage is the stage of giving labels to data that has gone through the preprocessing stage [24]. This research tries to utilize automatic labeling by utilizing the VADER library. VADER is an abbreviation of Valence Aware Dictionary and Sentiment Reasoner, which is a lexicon or dictionary and rule-based sentiment analysis tool that specifically labels expressions on social media, and also works well on text from other domains [25]. The following table takes an example of a dataset from table 2. Table 3 of the labeling results will be used in the next process, namely in creating a classification model.

Table 3. Dataset Labelling Using Vader

Document	Text	Class
D1	“lanjut” “diam” “erti” “tidak” “semua” “sama” “harus” “libat” “hati”	-1
D2	“lagu” “bosan”	-1
D3	“makasih” “hasil”	1
D4	“nomor” “satu” “bukan”	1

3.4 Feature Extraction

TF-IDF (Term Frequency-Inverse Document Frequency) is a feature extraction technique that uses statistics on word occurrences in documents. The goal of feature extraction is to convert data that is initially in text form into a numerical representation. Computers can only process and understand data in numerical form, therefore, data that was originally in the form of text, images or videos must be converted into numerical form before it can be processed further by the computer. This process is what allows computers to analyze and work with data that previously could not be understood directly by machines [26].

3.5 TF-IDF

At the extraction or weighting stage, a term separation process will be carried out as in the preprocessing stage, the difference is that in feature extraction there are several methods called n-grams. N-grams can be done in ways such as: n=1, n=2, n=3. The example of data preprocessing

 ISSN:2302-8149
e-ISSN:2540-9719

results in table 2 is used as an example of weighting calculations using TF-IDF. The following table applies n = 1, each term in the previous D1 will be separated and made into one corpus. At this time, it is not for example T1, where T represents the token and 1 represents the first term. The number then increases until the last term in the corpus. Dataset labelled using vader can be seen in table 4.

Table 4. Dataset Labelling Using Vader

Token	Term
T1	lanjut
T2	diam
T3	erti
T4	tidak
T5	semua
T6	sama
T7	harus
T8	libat
T9	hati

Each word in table 4 is calculated using formula 1.

$$W_{t,d} = 1 * \log(\frac{N}{dft}) \quad (1)$$

Taking the word "continue" as an example, the word "continue" in D1 contains 1. Then each document is checked, then the number of documents containing the searched word is added up, namely df, t =1. N= 4 is the total number of documents in the example. Then the Wt, d calculation is carried out using the following formula:

$$W_{t,d} = 1 * \log(\frac{4}{1}) = 0.602 \quad (2)$$

The recapitulation results can be seen in table 5 below.

Table 5. Recapitulation of Weight Calculations For Every Term

TF					DF	$IDF = (\log(\frac{N}{dft}))$	$W = tf_{t,d} * \log(\frac{N}{dft})$			
Term	D1	D2	D3	D4			D1	D2	D3	D4
lanjut	1	0	0	0	1	0.602	0.602	0	0	0
diam	1	0	0	0	1	0.602	0.602	0	0	0
erti	1	0	0	0	1	0.602	0.602	0	0	0
tidak	1	0	0	0	1	0.602	0.602	0	0	0
semua	1	0	0	0	1	0.602	0.602	0	0	0
sama	1	0	0	0	1	0.602	0.602	0	0	0
harus	1	0	0	0	1	0.602	0.602	0	0	0
libat	1	0	0	0	1	0.602	0.602	0	0	0
hati	1	0	0	0	1	0.602	0.602	0	0	0
Jumlah							5.418	0	0	0

3.6 Feature Selection

To form a population using swarm particles, first initialize the parameters. The initialized parameters are c, namely velocity control, p population size. The next step is to turn all the data into particles. Each particle calculates a fitness function. These fitness function values are then compared to determine the particle as Pbest or Gbest. The process repeats for a predetermined number of iterations. If the iteration is complete, the best parameters are selected [27]. In this research, Binary Particle Swarm Optimization (BPSO) is used, which is a variation of PSO [28]. The flow of PSO can be seen in Figure 3. BPSO will select influential features after the feature extraction process, which has been given a weight and obtained the number of features. Feature size is the dimension of particle search. These particles are all terms/words which are also called features. Each term is the initial position of the particle. The position of each particle is represented by binary values, namely 0 and 1. A value of 0 means the feature is not used and 1 means the feature is used at the next stage. The BPSO

 ISSN:2302-8149
e-ISSN:2540-9719

implementation uses pyswarm, a library provided by python. The parameters for feature selection using pyswarms are c1, c2, w, k, and p. c1 is a cognitive parameter, used is 0.5. c2 is a social parameter, which is used at 0.5. w is the inertia parameter, which is used at 0.9. k is the number of neighbors considered. This parameter value must be smaller than n_particles, which is 30. The p used is 2, meaning the distance calculation uses l2-norm or Euclidean norm.

3.7 Classification Model

The research sentiment analysis model is formed into two types, namely: (1) SVM model without PSO feature selection, (2) SVM model using PSO feature selection. The training process utilizes GridSearchCV, which is a method of tuning/searching for values in depth so that it tries all possible combinations of the desired parameters and finds the best one. GridSearchCV makes it easy to tune parameters in linear, poly-nomial and radial basis function (RBF) kernels. Each combination of experiments is trained using cross validation 5 times. From the 5 experiments, the average accuracy will be taken and the parameter combination with the greatest accuracy will be used. The application of the GridsearchCV method utilizes the library provided by Scikit-Learn. The second model is SVM combined with Particle Swarm Optimization feature selection.

3.8 Model Evaluation

The SVM classification model that has been formed will be evaluated using a confusion matrix. Confusion Matrix includes techniques including accuracy, f-1 macro and cross validation. Accuracy and f-1 macros are used to evaluate training data.

3.9 Analysis of Research Results

At this stage, an analysis of the model training results is carried out. In the first model, hyperparameter tuning uses the GridSearchCV method with parameter values that have been determined for each kernel. Each parameter combination was trained 5 times. Data is divided into training data and testing data. Training data is data used to train the model during the learning process. Meanwhile, testing data is data used to test models obtained from the learning process. The ratio of training data is generally greater than test data. Data is divided into a ratio of 80% and 20%. The labeling used is English-based VADER so that data from both sources is translated into English. Flowchart of PSO method can be seen in figure 3.

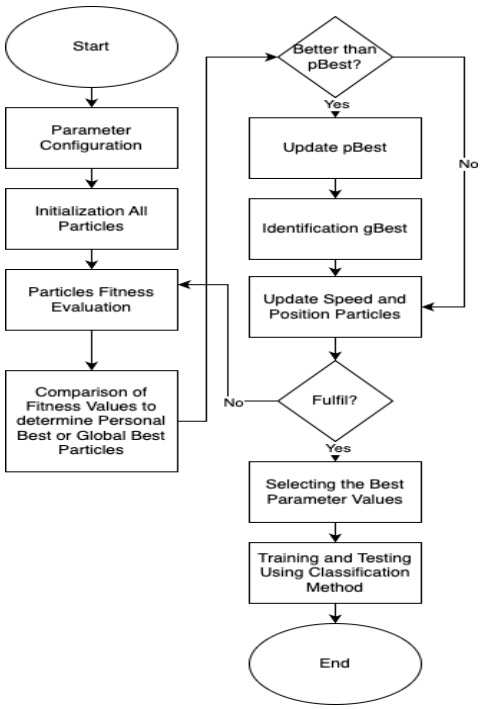

Figure 3. PSO method flowchart

 ISSN:2302-8149
e-ISSN:2540-9719

4 Result and Analysis

The stages in results and analysis include data labeling, feature extraction, data splitting, classification, evaluation of classification models, and analysis of research results.

4.1 Dataset Labelling

In this research, labeling uses VADER, a library provided by NLTK. VADER (Valance Aware Dictionary and Sentiment Reasoner) is a rule-based lexicon and sentiment analysis tool that is adapted to sentiments expressed on social media. Vader uses a combination of sentiment lexicons such as lexical features, generally labeled based on semantics. Vader will give a score to the document using the polarity method, giving a positive, negative and neutral polarity score. To categorize polarity into a compound score, a compound score is available. Compound score is a score calculated by averaging the valence score of each word based on the lexicon which is adjusted to the rules and normalized between the values -1 (meaning the most extreme negative) and +1 (meaning the most extreme positive). Then, this compound score will be labeled positive with a compound score >= 0.5, neutral if the compound score is between -0.5 and <0.5, negative if the compound score is <= -0.5. The results can be seen in figure 4.

The labeling scenario is carried out in two ways, namely: (1) The data is not translated into English and (2) the data is translated first into English. Scenario one with Indonesian language data labeled using Vader resulted in unbalanced data labels between negative, positive and neutral. Neutral data is more dominant, causing imbalanced data. Therefore, the second scenario was carried out, namely changing the data into English using the GoogleTrans library and then labeling it with Vader [29]. The comparison result of twitter data labels after labelling using vader can be seen in table 6.

Table 6. Comparison of Twitter Data Labels After Labelling Using Vader

Label	Label Using One hot encoding	Indonesian Data	Translation Data
Negative	0	22	165
Neutral	1	376	162
Positif	2	59	130
Total		**457**	**457**

The results of the polarity score snapshot after cleaning and tokenizing can be seen in Figure 4.

text	clean	tokens	score
_Assalamu?alaikum Wr Wb._\n\nAlhamdulillah tel...	a alamu alaikum wr wb alhamdulillah telah terb...	alhamdulillah terbit tabloid media umat edisi ...	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
gue happy bgt ada UU permendikbud pasal 30 aya...	kamu happy bgt ada uu permendikbud pasal ayat ...	happy bgt uu permendikbud pasal ayat barusan l...	{'neg': 0.0, 'neu': 0.893, 'pos': 0.107, 'comp...
Predator seks harus dibuang jauh dari kampus. ...	predator seks harus dibuang jauh dari kampus s...	predator seks buang kampus henti lindung balik...	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
Permendikbud-Ristek PPKS menjelaskan langkah a...	permendikbud ristek ppks menjelaskan langkah a...	permendikbud ristek ppks langkah cipta ruang a...	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
Istri Gubernur itu Ex-Officio Bunda PAUD berda...	istri gubernur itu ex officio bunda paud berda...	istri gubernur ex officio bunda paud dasar per...	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
...	...	...	...
@NayanikaPutra Permendikbud bs kuat krn bukan ...	permendikbud bs kuat karena bukan cuma didukun...	permendikbud bs kuat dukung sm dukung perintah...	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
Survei SMRC: Mayoritas Publik Dukung Permendik...	survei smrc mayoritas publik dukung permendikb...	survei smrc mayoritas publik dukung permendikb...	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
btw yang aku maksud permendikbud no. 30 tahun ...	btw yang aku maksud permendikbud no tahun ya j...	btw maksud permendikbud mi informasi	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
Masalahnya permendikbud tupoksinya apa, lahh d...	masalahnya permendikbud tupoksinya apa lahh di...	permendikbud tupoksinya lahh mi tupoksi ranah ...	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...
"Permendikbud tuh ngegiring opini orang kalo a...	permendikbud tuh ngegiring opini orang kalau a...	permendikbud ngegiring opini orang consent uda...	{'neg': 0.0, 'neu': 0.909, 'pos': 0.091, 'comp...

Figure 4. Results of giving polarity scores

4.2 Feature Extraction

After going through the preprocessing stage, this means that the data is already in word tokens. The next stage is to assign a weight to each previous token in the term-document matrix. The method used is TF-IDF, in the Python programming language implemented by the TfidfVectorizer library provided

 ISSN:2302-8149
e-ISSN:2540-9719

by Scikit learn. The number 0 represents the index, the number 807 represents the order of the terms in the corpus which have been sorted alphabetically and 0.210759 is the weight of the term. Figure 5 shown snippet of the appearance of the term in the document and figure six is the weight result from TF-IDF.

```
  (0, 807)      0.21075903959700626
  (0, 453)      0.28308024646539437
  (0, 661)      0.28775880506460033
  (0, 977)      0.29871241131108744
  (0, 86)       0.18755997066866503
  (0, 806)      0.17095233118218559
  (0, 452)      0.23383044210385456
  (0, 640)      0.051539688250274406
  (0, 976)      0.27124215302102905
  (0, 879)      0.29293071164960227
  (0, 47)       0.3128340658570126
  (0, 958)      0.28775880506460033
  (0, 522)      0.30526716177436447
  (0, 887)      0.22488069414293135
  (0, 29)       0.3128340658570126
```

Figure 5. Snippet of the appearance of the term in the document

	abis	ada	adil	adlh	adu	advokat	advokat harap	agama	agama sarang	agenda	...	wib link	wkwkwk	you	youtube	yuk	zina	zina generasi	zina joko	zina zina	zinah
D1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.00000	0.000000	0.0	0.0	0.0	0.0
D2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.00000	0.000000	0.0	0.0	0.0	0.0
D3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.00000	0.244082	0.0	0.0	0.0	0.0
D4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.28617	0.000000	0.0	0.0	0.0	0.0
D5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.00000	0.000000	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
D1288	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.00000	0.000000	0.0	0.0	0.0	0.0
D1289	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.00000	0.000000	0.0	0.0	0.0	0.0
D1290	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.00000	0.000000	0.0	0.0	0.0	0.0
D1291	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.00000	0.235141	0.0	0.0	0.0	0.0
D1292	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.00000	0.244981	0.0	0.0	0.0	0.0

Figure 6. Dataset that has been weighted using TF-IDF

Pay attention to Figure 6, there are rows and columns where the rows state the number of documents and the columns are the terms that exist in the entire dataset, while the values in the table are the TF IDF values of the terms (columns) that are obtained if they are present in each row, if is not there then the value will be 0.0. The results of the stages are used as features in the next process.

4.3 Feature Selection

The feature selection process uses the psywarm library with variations of binary particle swarm optimization (BPSO). This library is provided by Python which requires input parameters including the number of swarm sizes, search dimensions, inertia weight, acceleration coefficient and maximum particle neighbors. Each particle evaluates its fitness value based on the accuracy value of the subset features being tested. The best position of each particle will be known at the end of the iteration. The BPSO parameters used were taken from previous research, namely c1 = 0.5, c2 = 0.5, w = 0.9, k = 30, and p = 2. PSO experiment with 100 iterations.

4.4 Classification Model

The SVM parameter in the linear kernel used is C or cost to form a curve from the hyperplane, curve means how the hyperplane separates the data. The parameter in the RBF kernel that has an influence is C or cost, gamma (γ) to maximize margin. The parameters that influence the polynomial kernel are: C and degree. The degree parameter is a separating function in the polynomial kernel. In this study, gamma, linear and polynomial kernels were used with gamma and C parameters of 0.01, 0.1, 1, 10, 100, d = 1, 2 specifically for polynomial kernels. Gridsearch is a method for selecting a combination of models and hyperparameters. This method tests combinations one by one and determines the combination with the best model performance which is then used as a model for prediction. The following is a comparison of accuracy using gridsearch. The results of tuning parameters with the best

 ISSN:2302-8149
e-ISSN:2540-9719

accuracy will be used for the testing process on each kernel. The following is a summary of the parameters used for the next stage which can be seen in table 10 for the Twitter dataset .

Table 10. Initial Parameter In Twitter Dataset

Data/Parameter	Kernel Initialization	Data Twitter (Indonesia)	Kernel Initialization	Translation Data
Kernel Linear	K11	C = 10, *gamma* = 0.01	K12	C = 1, *gamma* = 0.01
Kernel RBF	R11	C = 10, *gamma* = 0.01	R12	C = 100, *gamma* = 0.01
Kernel Polynomial	P11	C = 100, *gamma* = 0.1, d=1	P13	C = 10, *gamma* =0.1, d =1
Kernel Polynomial	P12	C = 10, *gamma* = 1, d =2	P14	C = 1, *gamma* = 10, d = 2

4.5 Confusion Matrix in Models Without Particle Swarm Optimization

Accuracy of Indonesian Language Twitter Data Using PSO can be seen in table 11.

Table 11. Accuracy of The Indonesian Twitter Dataset Model After Using PSO

No	Nama Kernel	Kernel Initialization	PSO Parameter
Linear	L11	c1=0.5, c2=0.5,	0.85
RBF	R11	w=0.9, k=30, p=2	0.83
Polynomial	P11		0.87
Polynomial	P12		0.83

Table 11 explains the performance results of the model after PSO feature selection was carried out on the Twitter dataset which was translated into English for each kernel with the parameters obtained when tuning the hyperparameters.

4.6 Accuracy of Twitter Data Translated to English Using PSO

Accuracy of English Language Twitter Data Using PSO can be seen in table 12.

Table 12. Model Accuracy on A Twitter Dataset Translated Into English Using PSO

	Nama Kernel	Kernel Initialization	PSO Parameter
Linear	L112	c1=0.5, c2=0.5,	0.55
RBF	R12	w=0.9, k=30, p=2	0.45
Polynomial	P13		0.51
Polynomial	P14		0.34

Table 12 explains the performance results of the model after PSO feature selection. on the Twitter dataset which is translated into English in each kernel with the parameters obtained when tuning the hyperparameters.

4.7 Analysis of Model Performance Results

A comparison of research results is shown in the following table. The following table contains kernel initials, these kernel initials refer to table 13.

 ISSN:2302-8149
e-ISSN:2540-9719

Table 13. Comparison of Model Accuracy Before and After Using PSO

Data / Algorithm	SVM without PSO				SVM Using PSO			
	Accuracy				Accuracy			
Kernel Initials	L11	R1	P11	P12	L12	R2	P21	P22
PSO					C1 = 0.5, c2=0.5, w=0.9, k=30, p=2, iterasi = 100			
Data/Parameters								
Indonesian Twitter Data	**0.87**	**0.76**	0.86	**0,78**	0.85	**0.83**	0.87	**0.83**
Twitter Data Translated into English	**0.61**	0.60	0.60	0.29	**0.55**	0.45	0.51	0.34

Table 13 is a summary of model accuracy. The accuracy of the Indonesian language Twitter dataset model was compared with the dataset after being translated into English. This is to see the performance of VADER automatic labeling. Then, the accuracy in the two previous scenarios is compared after applying PSO feature selection. This is to measure whether PSO has succeeded in selecting important features. In the model with the Twitter dataset, the model with linear kernel, RBF, polynomial order = 1, polynomial order 2, obtained the highest accuracy on the Indonesian language dataset with values of 0.87, 0.76, 0.86, 0.78 respectively. Linear Kernel managed to get the highest accuracy. Meanwhile, after the dataset was translated, the accuracy was sequentially smaller, linear kernel, RBF, 1st order polynomial, 2nd order polynomial, 0.61, 0.60, 0.60, 0.29. This is because the English language dataset does not go through the same preprocessing stages as the Indonesian language dataset. The application of feature selection is expected to increase accuracy, but PSO was not successful in increasing accuracy for each kernel. In the Indonesian language Twitter dataset, accuracy increases in the RBF kernel and 2nd order polynomial. On the Twitter dataset, after being translated, feature selection was not successful in all kernels. The accuracy obtained actually decreases. The result of comparison F1 macro model can be seen in table 14.

Table 14. F1 Macro Model Comparison

Data/Algorithm	SVM Without PSO			
	F-1 Macro			
Kernel Initials	**L11**	**R1**	**P11**	**P12**
PSO Data/Parameter				
Twitter data in Indonesian	**0.71**	0.29	0.68	**0.41**

Table 14 explains the results of the f1-macro model for each kernel. Measuring f1-macro due to the imbalanced condition of the dataset. Based on the f1-macro value, the 2nd order polynomial kernel produces similar performance, namely linear kernel, RBF, 1st order polynomial and 2nd order polynomial of 0.87, 0.87, 0.80, 0.80 on the Indonesian language Twitter dataset.

Based on the results of accuracy and f1-macro on the model before applying PSO. In the Indonesian language Twitter dataset, the best performance is the linear kernel. In the Dataset after being translated the best performance is achieved by the linear kernel.

5 Conclusion

Based on the research findings, it can be concluded that the model evaluation using the confusion matrix yielded different performance results. The metrics used were accuracy and f1-macro. In both the Indonesian language Twitter dataset and the English-translated Twitter dataset, it was observed that the linear kernel performed the best. For the Indonesian language Twitter dataset, the accuracy

 ISSN:2302-8149
e-ISSN:2540-9719

was 0.87 with an f1-macro of 0.71. Meanwhile, for the English-translated Twitter dataset, the accuracy was 0.61 with an f1-macro of 0.59.
Furthermore, the model evaluation after applying Particle Swarm Optimization (PSO) feature selection resulted in varying accuracies. In the Indonesian language Twitter dataset, kernels such as RBF kernel and first-order and second-order polynomials succeeded in increasing accuracy. The accuracy improved from 0.76 before PSO to 0.78 to 0.82 and 0.81 after PSO. However, on the English-translated Twitter dataset, the feature selection failed to increase accuracy.

References

[1] A. R. M. A. Ramdhani, "Konsep Umum Pelaksanaan Kebijakan Publik," *J. Publik*, vol. Vol 11, no. January, pp. 1–12, 2016, [Online]. Available: https://journal.uniga.ac.id/index.php/JPB/article/download/1/1.

[2] R. N. Andari and P. Kajian, "Evaluasi Kebijakan Penanganan Kejahatan Kekerasan Sesksual Terhadap Anak di Indonesia (Evaluation Policy of Carrying Out of Sexual Violence Crimes of Children)," *J. Ilm. Kebijak. Huk.*, vol. 11, no. 1, pp. 1–11, 2017, [Online]. Available: http://ejournal.balitbangham.go.id/index.php/kebijakan/article/view/86/pdf_1.

[3] F. Y. Hardianti, R. Efendi, P. D. Lestari, and E. S. Puspoayu, "Urgensi Percepatan Pengesahan Rancangan Undang-Undang Penghapusan Kekerasan Seksual," *J. Suara Huk.*, vol. 3, no. 1, p. 26, 2021, doi: 10.26740/jsh.v3n1.p26-52.

[4] B. Liu, *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. 2015.

[5] I. F. Ashari, "Analysis Sentiments In Facebook Down Case Using Vader And Naive Bayes Classification Method," *Multitek Indones. J. Ilm.*, vol. 16, no. 2, pp. 75–89, 2022.

[6] N. Mtetwa, A. O. Awukam, and M. Yousefi, "Feature Extraction and Classification of Movie Reviews," *5th Int. Conf. Soft Comput. Mach. Intell. ISCMI 2018*, pp. 67–71, 2018, doi: 10.1109/ISCMI.2018.8703235.

[7] Y. Wang, Z. Zhou, S. Jin, D. Liu, and M. Lu, "Comparisons and Selections of Features and Classifiers for Short Text Classification," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 261, no. 1, 2017, doi: 10.1088/1757-899X/261/1/012018.

[8] V. Mickevičius, T. Krilavičius, and V. Morkevičius, "Classification of Short Legal Lithuanian Texts," pp. 10–11, 2015, [Online]. Available: http://www.agendasetting.dk.

[9] I. F. Ashari, F. A, Daffa, and S. A, "Sentiment Analysis of Tweets About Allowing Outdoor Mask Wear Using Naïve Bayes and TextBlob," *Indones. J. Comput. Sci.*, vol. 12, no. 3, pp. 1092–1103, 2023, doi: 10.33022/ijcs.v12i3.3238.

[10] M. D. Alizah, A. Nugroho, U. Radiyah, and W. Gata, "Sentimen Analisis Terkait 'Lockdown' pada Sosial Media Twitter," *CSRID (Computer Sci. Res. Its Dev. Journal)*, vol. 12, no. 3, p. 143, 2021, doi: 10.22303/csrid.12.3.2020.143-149.

[11] T. Mardiana, H. Syahreva, and T. Tuslaela, "Komparasi Metode Klasifikasi Pada Analisis Sentimen Usaha Waralaba Berdasarkan Data Twitter," *J. Pilar Nusa Mandiri*, vol. 15, no. 2, pp. 267–274, 2019, doi: 10.33480/pilar.v15i2.752.

[12] S. W. Yudha and M. Wahyudi, "Komparasi Algoritma Klasifikasi Untuk Analisis Sentimen Review Film Berbahasa Asing," *Semin. Nas. Inform. Sist. Inf. Dan Keamanan Siber*, pp. 180–185, 2018.

[13] S. C. Rachiraju and M. Revanth, "Feature Extraction and Classification of Movie Reviews using Advanced Machine Learning Models," *Proc. Int. Conf. Intell. Comput. Control Syst. ICICCS 2020*, no. Iciccs, pp. 814–817, 2020, doi: 10.1109/ICICCS48265.2020.9120919.

[14] P. M. R. C. Dinatha and N. A. Rakhmawati, "Komparasi Term Weighting dan Word Embedding pada Klasifikasi Tweet Pemerintah Daerah," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 9, no. 2, pp. 155–161, 2020, doi: 10.22146/jnteti.v9i2.90.

[15] M. C. Untoro, M. Praseptiawan, I. F. Ashari, and A. Afriansyah, "Evaluation of Decision Tree, K-NN, Naive Bayes and SVM with MWMOTE on UCI Dataset," *J. Phys. Conf. Ser.*, vol. 1477, no. 3, 2020, doi: 10.1088/1742-6596/1477/3/032005.

[16] L. B. Ilmawan and M. A. Mude, "Perbandingan Metode Klasifikasi Support Vector Machine dan Naïve Bayes untuk Analisis Sentimen pada Ulasan Tekstual di Google Play Store," *Ilk. J. Ilm.*, vol. 12, no. 2, pp. 154–161, 2020, doi: 10.33096/ilkom.v12i2.597.154-161.

[17] I. Subagyo, L. D. Yulianto, W. Permadi, and A. W. Dewantara, "Sentiment Analisis Review Film Di IMDB Menggunakan Algoritma SVM," *e-Jurnal JUSITI (Jurnal Sist. Inf. dan Teknol. Informasi)*, vol. 8–1, no. 1, pp. 47–56, 2019, doi: 10.36774/jusiti.v8i1.600.

[18] I. Rish, "An Empirical Study of the Naïve Bayes Classifier An empirical study of the naive Bayes classifier," *Cc.Gatech.Edu*, no. January 2001, pp. 41–46, 2014, [Online]. Available: https://www.cc.gatech.edu/~isbell/reading/papers/Rish.pdf.

[19] I. F. Ashari, M. C. Untoro, E. Maharani Sutowo, D. Salsabila, and D. Athiyah Zhabiyan, "Hyperparameter Tuning Feature Selection with Genetic Algorithm and Gaussian Naïve Bayes for Diabetes Disease Prediction," *J. Telemat.*, vol. 17, no. 1, 2022, [Online]. Available: https://www.researchgate.net/publication/365036216.

[20] Y. Lu, M. Liang, Z. Ye, and L. Cao, "Improved particle swarm optimization algorithm and its application in text feature selection," *Appl. Soft Comput. J.*, vol. 35, pp. 629–636, 2015, doi: 10.1016/j.asoc.2015.07.005.

[21] S. Ernawati, R. Wati, N. Nuris, L. S. Marita, and E. R. Yulia, "Comparison of Naïve Bayes Algorithm with Genetic Algorithm and Particle Swarm Optimization as Feature Selection for Sentiment Analysis Review of Digital Learning Application," *J. Phys. Conf. Ser.*, vol. 1641, no. 1, 2020, doi: 10.1088/1742-6596/1641/1/012040.

[22] I. Syarif, "Feature Selection of Network Intrusion Data using Genetic Algorithm and Particle Swarm Optimization," *Emit. Int. J. Eng. Technol.*, vol. 4, no. 2, pp. 277–290, 2016, doi: 10.24003/emitter.v4i2.149.

[23] Muttaqin, Yuswardi, A. Maulidinnawati, A. Parewe, I. F. Ashari, and M. Munsarif, *Pengantar Sistem Cerdas*. 2023.

[24] Muttaqin *et al.*, *Data Mining Teori dan Implementasi*, vol. 1. 2023.

[25] E. Hutto, C.J. and Gilbert, "VADER: A Parsimonious Rule-based Model for," *Eighth Int. AAAI Conf. Weblogs Soc. Media*, p. 18, 2014, [Online]. Available: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/viewPaper/8109.

[26] Z. Rustam, E. Sudarsono, and D. Sarwinda, "Random-Forest (RF) and Support Vector Machine (SVM) Implementation for Analysis of Gene Expression Data in Chronic Kidney Disease (CKD)," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 546, no. 5, 2019, doi: 10.1088/1757-899X/546/5/052066.

[27] I. Cholissodin and E. Riyandani, "Buku Ajar Swarm Intelligence," no. June, pp. 1–198, 2016.

[28] L. Shang, Z. Zhou, and X. Liu, "Particle swarm optimization-based feature selection in sentiment classification," *Soft Comput.*, vol. 20, no. 10, pp. 3821–3834, 2016, doi: 10.1007/s00500-016-2093-2.

[29] B. A. Prasetyo, "Analisis Sentimen Pengguna Twitter untuk Text Berbahasa Indonesia terhadap Pelayanan Home Fix Broadvand," *Anal. Sentimen Pengguna Twitter untuk Text Berbahasa Indones. terhadap Pelayanan Home Fix Broadvand*, no. September, pp. 18–23, 2021.