BAREKENG: Journal of Mathematics and Its Applications
March 2023 Volume 17 Issue 1 Page 0481–0486
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol17iss1pp0481-0486

SENTIMENT ANALYSIS OF MERDEKA BELAJAR KAMPUS MERDEKA POLICY USING SUPPORT VECTOR MACHINE WITH WORD2VEC

Nurul Rezki [1], Sri Astuti Thamrin[2*], Siswanto[3]

[1]PT. Kioser Teknologi Indonesia
Jl. Puri Asri Raya Tamalanrea Indah, Makassar, 90245, Indonesia

[2,3]Department of Statistics, Faculty of Mathematics and Natural Sciences, University of Hasanuddin
Jl. Perintis Kemerdekaan. KM 10, Makassar, 90245, Indonesia

*Corresponding author's e-mail: * **tuti@unhas.ac.id***

ABSTRACT

Article History:

Received: 16th November 2022
Revised: 3rd February 2023
Accepted: 14th February 2023

Keywords:

Merdeka Belajar Kampus Merdeka Policy;
Sentiment Analysis;
Support Vector Machine;
Twitter;
Word2Vec

Sentiment analysis is a data text analysis that classifies data into positive and negative sentiments. This study aims to obtain the results of sentiment classification related to Merdeka Belajar Kampus Merdeka policy on Twitter using a support vector machine algorithm with Word2Vec feature extraction. Support Vector Machine is a classification algorithm that separates data classes using the optimum hyperplane. Text data in sentiment analysis must change its numerical form by performing feature extraction. In this study, the feature extraction used is Word2Vec which represents words in vector form. Data in this study are tweets with the keyword "Kampus Merdeka" uploaded on Twitter as many as 10000 tweets. After preprocessing text data, data used to analyze sentiment was 1579 tweets. Sentiment classification resulted in a classification model accuracy of 89.87%, a precision of 91.20%, a recall of 84.44%, and F-Measure of 87.68%. Classification sentiment using a support vector machine with Word2Vec feature extraction in this study produces a good model.

How to cite this article:

N. Rezki, S. A. Thamrin and Siswanto., "SENTIMENT ANALYSIS OF MERDEKA BELAJAR KAMPUS MERDEKA POLICY USING SUPPORT VECTOR MACHINE WITH WORD2VEC," *BAREKENG: J. Math. & App.*, vol. 17, iss. 1, pp. 0481-0486, March 2023.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article • Open Access

1. INTRODUCTION

Sentiment is a person's opinion about feelings, attitudes, or thoughts that can. Individual sentiment towards a particular event, brand, product or company can be obtained from news reports, user reviews, social media, or microblogging sites [1]. Sentiment-related data obtained through websites or social media are generally in the form of text and unstructured.

Text mining is a pattern extraction process from some unstructured data and will obtain data on patterns, trends, and potential extraction from text data [2]. One of the purposes of using text mining is sentiment analysis, namely the process of understanding, extracting, and processing textual data to obtain sentiment information contained in a sentence of opinion on a problem or object. Sentiment information is obtained as positive or negative sentiment [3] [4].

There are several general algorithms that can be used to perform sentiment analysis, such as Support Vector Machine (SVM), Naive Bayes, and Maximum Entropy. SVM is one of the algorithms that is widely used in studies related to sentiment analysis because it produces good accuracy compared to other algorithms or classification methods. One of them is a sentiment analysis of an airline's review using the supporting vector machine method and Naive Bayes, which produces the highest accuracy of 82.48% for the supporting vector machine method [5].

Sentiment data in the form of text data which is analyzed by sentiment analysis needs to be converted into the numeric form so that it can be read by a computer. Converting text data into numeric data can be done by feature extraction [6]. There are several types of feature extraction, such as Term Frequency (TF), Term Frequency Inverse Document Frequency (TF-IDF), Word2Vec, Glove, and so on [7]. In this study, the feature extraction used is Word2Vec which represents a word in vector form [8]. In another study of sentiment analysis on Twitter related to public policy, compared to TF-IDF, Word2Vec feature extraction provides increased accuracy in the SVM method [9].

One of the issues currently being discussed on various social media and websites is the Merdeka Belajar Kampus Merdeka (MBKM) policy by the Ministry of Education and Culture of the Republic of Indonesia. Based on an article released on January 29, 2020, by tirto.id, since its initial launch, this program has received many pro and contra reactions from the public [10]. The reaction was conveyed through social media such as Twitter, Instagram, Facebook, and so on. Twitter is one of the most popular social media in Indonesia. In addition, Twitter provides an Application Programming Interface (API) for users to develop an application with data sourced from Twitter [11].

2. RESEARCH METHODS

Data used in this study is primary data obtained from Twitter in the form of Indonesian-language tweets with the keyword "Kampus Merdeka" uploaded on January 20, 2020, to March 31, 2022. 10,000 tweet data used is in the form of text.

The data structure used in this study after preprocessing the text of the tweet data consists of predictor variables, namely the basic words of each tweet, and the response variable, namely the classification of tweet sentiments (positive and negative). The data classification in the supporting vector method is divided into training data and testing data. Table 1 shows an example of research structure data before preprocessing.

Table 1. Research Data Structure

No	Tweet	Sentiment
1	@Batutuo terus berjuang wujudkan kampus merdeka ...disain program2nya agar dihasilkan SDM yg berkwalitas di Era 4.0	Positive
2	Program Kampus Merdeka ini keren banget sih, ngasih harapan buat perubahan sistem pendidikan Indonesia	Positive
3	Ngiri saya sama program kampus merdeka	Positive
⋮	⋮	⋮
1579	labelnya kampus merdeka, tapi sesungguhnya sivitasnya terjajahnya sama berlapis-lapis regulasi	Negative

The stages of analysis in this study are as follows:

1. Crawling tweet data using snscrape, which is saved in csv format.
2. Perform manual labeling using Doccano, which is an open-source tool for data annotation.

3. Perform sentiment analysis with the supporting vector machine method while the stages of analysis are as follows:
   A. For data training:
      a. Preprocess data.
      b. Perform feature extraction Word2Vec with Skip-Gram architecture and negative sampling algorithm.
      c. Perform sentiment analysis with the SVM method.
      d. Obtaining a classification model.
   B. For data testing:
      a. Preprocess data.
      b. Perform feature extraction Word2Vec with Skip-Gram architecture and negative sampling algorithm.
      c. Perform sentiment analysis with the SVM method.
      d. Get sentiment analysis predictions.
4. Evaluation of SVM classification model using Word2Vec feature extraction with a confusion matrix.

3. RESULTS AND DISCUSSION

3.1. Descriptive Analysis

Data was obtained from crawling on Twitter with the keyword "Kampus Merdeka" which was uploaded on January 20, 2020 to March 30, 2022. The crawling results obtained as many as 10000 Indonesian-language tweets. Results of data collection are obtained in Table 2.

Table 2. Research Data Structure

No	Publication Date	Tweet	Username
1	2020-02-20 07:50:02+00:00	[MAU TAHU] Luncurkan Empat Kebijakan Merdeka Belajar, Menteri @Kemdikbud_RI Beri Tajuk Kampus Merdeka https://t.co/0ZI4ny2O3g #BersamaIndonesiaMaju #SDMUnggulIndonesiaMaju #JokowiMajukanPendidikan https://t.co/G3gaGQLrcT	tigapilarnews
2	2020-02-08 10:15:28+00:00	@Kemdikbud_RI. Pak Nadiem Makarim; Kampus Merdeka #MerdekaBelajar juga harus #MerdekaDariKekerasan seksual. Segera terbitkan SOP penanganan &amp; pencegahan kekerasan seksual di instansi pendidikan. - Tandatangani Petisi! https://t.co/4MR1BB119v lewat @ChangeOrg_ID	e100ss
3	2020-02-09 04:26:27+00:00	@adkestwit_ugm Kampus Merdeka bikin saya merdeka dari bayar UKT ga min?	AkuSukaSambal
⋮	⋮	⋮	⋮
10000	2022-03-31 13:32:03+00:00	Btw Kampus Merdeka ada banyak banget programnya. Ada proyek kemanusiaan, pertukaran mahasiswa, magang, riset atau penelitian, mengajar, wirausaha. Malah katanya mau ada program Bangkit juga tuh program kesiapan kerja yg didesain Google. Mantap ya #G20BersamaIndonesia	mba_diahworo

Manual labeling was carried out on collected data. Data is labeled into two classes of sentiment, namely positive and negative. Positive sentiment about support for MBKM, while negative sentiment contains policy distrust towards MBKM. Based on the results of manual labeling of 10000 tweet data, 1579 data were labeled with data, and 8421 data were ignored because they were noise data. Noise data is data with a large amount of additional information that is impossible, for example, tweets whose sentences cannot be categorized into positive or negative sentiments. A comparison of the amount of labeled data and noise data is significant. This is illustrated in the time span of data collected. More tweets containing additional information that does not mean tweets containing problems related to MBKM policies so that they cannot be categorized as positive or negative sentiments. Figure 1 shows the Bar Chart of MBKM's policy tweet sentiment class.

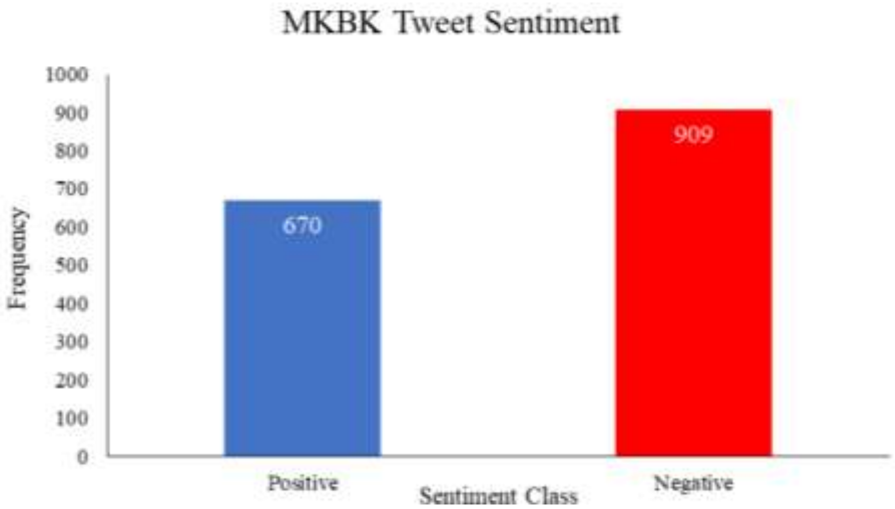

Figure 1. **MBKM Policy Data Sentiment Class Bar Chart**

Figure 1 has more negative sentiment data than positive sentiment data, it means that in the time span of the data collected, more people upload tweets on Twitter containing this matter compared to tweets containing support for MBKM policies.

3.2. Text Data Preprocess

Text data preprocessing is carried out on tweet data related to the collected MBKM policies. The preprocessing of the data text consists of data cleansing, case folding, spelling normalization, stemming, stopword removal, and tokenizing. Table 3 shows structure data before and after text data preprocessing.

Table 3. **Data Structure Before and After Text Data Preprocessing**

No	Data Structure Before Data Preprocessing	Data Structure After Data Preprocessing
1	Mantap.@nadiemmakarim @pusdatin_dikbud #KampusMerdeka https://t.co/H0VsvbV0Hr	[mantap]
2	"""@daastufff_ @collegemenfess Ikut program kampus merdeka aja kak, smt 3-5 boleh ambil prodi lain/study abroad dengan syarat dan ketentuan yang berlaku, kayak ada minimal IP"""	[ikut, program, kampus, merdeka, kakak, semester, ambil, prodi, lain, belajar, luar, negeri, syarat, tentu, laku, minimal, nilai]
3	Kampus Merdeka memberikan nilai plus pada kompetensi mahasiswa. https://t.co/2b64dwXXan	[kampus, merdeka, beri, nilai, tambah, kompetensi, mahasiswa]
⋮	⋮	⋮
1579	@TaheggaAlfath Menggaungkan kampus merdeka tapi kampus masih ndk boleh diskusi yang katanya bertema sensitif.	[gaung, kampus, merdeka, kampus, masih, tidak, boleh, diskusi, kata, tema, sensitif]

3.3. Word2Vec Feature Extraction

Word2Vec feature extraction is performed to obtain a vector of each base word from the tokenizing process [12]. Word vector is used as input data on the computer to build a SVM classification model. The basic words obtained from 1579 tweet data are 490 words. Table 4 shows the vectors of each base word obtained from the Word2Vec feature extraction process with the Skip-Gram architecture 100 dimensions $(h = 100)$ and window size 2 $(C = 2)$ with a negative sampling algorithm with 17 negative samples $(k = 17)$. Each word vector is 1×100 in size with vector elements in the form of numbers representing each basic word and are further used as data input in sentiment analysis with the SVM method.

Table 4. Basic Word Vector with Word2Vec Fitur Feature Extraction

No	Basic Word	dim1	dim2	dim3	dim4	⋯	dim100
1	merdeka	-0.15	0.46	0.07	0.05	⋯	0.02
2	kampus	-0.16	0.47	0.08	0.05	⋯	0.02
3	mahasiswa	-0.16	0.49	0.06	0.04	⋯	0.02
4	banget	-0.17	0.46	0.05	0.03	⋯	0.01
5	magang	-0.17	0.46	0.05	0.04	⋯	0.02
⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮
488	beban	-0.60	0.45	0.06	0.03	⋯	0.02
489	Kejar	-0.15	0.46	0.08	0.03	⋯	0.01
490	zenius	-0.17	0.47	0.06	0.03	⋯	0.02

3.4. Support Vector Machine Classification

Data classification with SVM consists of data that can be separated linearly and non-linearly. A trial-and-error method was carried out to determine the best supporting vector machine classification model. The classification training model is carried out on data training and model validation using data testing, in this study using training data and data testing with a ratio of 80:20. Using training data to build a classification model by the trial-and-error method. The best classification model is obtained, which is the vector engine supporting the RBF kernel, which is one of the kernels used for non-linearly separated data [13] [14].

Classification model training with a vector engine supporting the RBF kernel considers two parameters, namely C and determined through trial and error [15]. C is a parameter that functions to avoid misclassification of each sample tested, while it is a parameter that determines the support vector based on the effect of the sample distance with the decision limit. The value of parameter C is 1, and 0.2875 indicates the best model performance. Based on these parameters, the optimum hyperplane equation is obtained as follows:

$$f(x_j) = \sum_{i=1}^{n} \alpha_i y_i K\left(x_i, x_j\right) + b \tag{1}$$

$$= \textstyle\sum_{i=1}^{1579} \alpha_i y_i \exp\left(-\frac{(x_{\mathrm{i}}-x_{\mathrm{j}})^T (x_i - x_{\mathrm{j}})}{0.16}\right) + b$$

$$= \textstyle\sum_{i=1}^{1579} \alpha_i y_i \exp\left(-\frac{(x_{\mathrm{i}}-x_{\mathrm{j}})^T (x_i - x_{\mathrm{j}})}{0.16}\right) + \left(y_i - \frac{1}{2}\sum_{i=1}^{1579} \alpha_i y_i K(x_i, x_j)\right)$$

$$f(x_j) = \sum_{i=1}^{1579} \alpha_i y_i \exp\left(-\frac{(x_{\mathrm{i}} - x_{\mathrm{j}})^T (x_i - x_{\mathrm{j}})}{0.16}\right) + \left(y_i - \frac{1}{2}\sum_{i=1}^{1579} \alpha_i y_i \exp\left(-\frac{(x_{\mathrm{i}} - x_{\mathrm{j}})^T (x_i - x_{\mathrm{j}})}{0.16}\right)\right)$$

Validation of the classification model on data testing produces the confusion matrix in Table 5 below, which is used to measure the performance of the classification model.

Table 5. Confusion Matrix SVM Classification Model

Actual Class	Predict Class	
	Positive	Negative
Positive	114	21
Negative	11	170

Based on the Confucian matrix in Table 5, the accuracy of the classification model is 89.87%, namely the percentage of positive and negative sentiment data that is correctly predicted. The precision of 91.20% is the percentage of positive sentiment data that is correctly predicted to the overall data that is predicted to have a positive sentiment. The recall of 84.44% is a percentage, which is the percentage of positive sentiment data that is correctly predicted. Overall, the actual data with positive sentiment is positive. F-Measure of 87.68% shows the percentage of precision and recall simultaneously.

4. CONCLUSIONS

Merdeka Belajar Kampus Merdeka policy classification on Twitter uses a SVM with Radial Basis Function kernel and Word2Vec feature extraction resulting in a classification model accuracy of 89.87%, a precision of 91.20%, recall of 84.44%, and F-Measure 87.68%.

REFERENCES

[1] M. v. Mäntylä, D. Graziotin, and M. Kuutila, “The evolution of sentiment analysis - A review of research topics, venues, and top cited papers,” *Comput Sci Rev*, vol. 27, pp. 16–32, Feb. 2018.

[2] E. Turban, *Decision Support and Business Intelligence Systems*. London: Pearson Education, 2011.

[3] L. Zhang, S. Wang, and B. Liu, “Deep learning for sentiment analysis: A survey,” *WIREs Data Mining and Knowledge Discovery*, vol. 8, no. 4, Jul. 2018.

[4] B. Liu and L. Zhang, “A Survey of Opinion Mining and Sentiment Analysis,” in *Mining Text Data*, New York: Springer, 2012, pp. 415–463.

[5] A. M. Rahat, A. Kahir, and A. K. M. Masum, “Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset,” *International Conference on System Modeling and Advancement in Research Trends*, vol. 8, no. 2, pp. 266–270, Nov. 2020.

[6] A. M. Pravina, I. Cholissodin, and P. P. Adikara, “Analisis Sentimen Tentang Opini Maskapai Penerbangan pada Dokumen Twitter Menggunakan Algoritme Support Vector Machine (SVM),” *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 3, no. 3, pp. 2789–2797, Mar. 2019.

[7] M. Rusli, M. R. Faisal, and I. Budiman, “Ekstraksi Fitur Menggunakan Model Word2Vec untuk Analisis Sentimen pada Komentar Facebook,” *Seminar Nasional Ilmu Komputer (SOLITER)*, vol. 2, pp. 104–109, Oct. 2019.

[8] G. W. Aldiansyah, P. P. Adikara, and R. C. Wihandika, “Rekomendasi Lagu Cross Language Berdasarkan Lirik Menggunakan Word2VEC,” vol. 3, no. 8, pp. 8036–8041, Aug. 2019.

[9] H. F. Naufal and E. B. Setiawan, “Ekspansi Fitur Pada Analisis Sentimen Twitter Dengan Pendekatan Metode Word2Vec,” *e-Proceeding of Engineering*, vol. 8, no. 5, pp. 10339–10349, Oct. 2021.

[10] H. Prabowo, “Pro dan Kontra atas Kebijakan ‘Kampus Merdeka’ Nadiem,” *tirto.id*, Jan. 29, 2020, [Online]. Available: https://tirto.id/pro-dan-kontra-atas-kebijakan-kampus-merdeka-nadiem-evs2 (accessed Nov. 08, 2022).

[11] J. Blanchette, *The Little Manual of API Design*. Oslo: Trolltech, 2008.

[12] E. L. Goodman, C. Zimmerman, and C. Hudson, “Packet2Vec: Utilizing Word2Vec for Feature Extraction in Packet Data,” Apr. 2020.

[13] J. Cervantes, F. Gracia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, “A comprehensive survey on support vector machine classification: Applications, challenges and trends,” *Neurocomputing*, vol. 408, pp. 189–215, Sep. 2020.

[14] J. H. Jaman and R. Abdulrohman, “Sentiment Analysis of Customers on Utilizing Online Motorcycle Taxi Service at Twitter with The Support Vector Machine,” *International Conference on Electrical Engineering and Computer Science (ICECOS)*, pp. 231–234, Oct. 2019.

[15] S. Han, C. Qubo, and H. Meng, “Parameter selection in SVM with RBF kernel function,” in *Work Automation Congress 2012*, Jun. 2012, pp. 1–4.