Int. J. Nonlinear Anal. Appl. 13 (2022) No. 1, 1367–1373
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2021.5735

Sentiment analysis for covid-19 in Indonesia on Twitter with TF-IDF featured extraction and stochastic gradient descent

Vindi Dwi Antonio[a,*], Syahril Efendi[a], Herman Mawengkang[a]

[a] *Department of Master in Informatic Engineering, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia*

(Communicated by Madjid Eshaghi Gordji)

Abstract

Twitter is an information platform that can be used by any internet user. The opinions of the Twitter Netizens are still random or unclassified. The technique for classifying sentiment analysis requires an algorithm. One of the classification algorithms is Stochastic Gradient Descent (SGD). The more training data provided to the machine, the accuracy of the classification function model formed by the machine is also higher. But in making representations into numerical vectors, the dimensions of data become large due to the many features. Feature optimization needs to be done to the training data by reducing the dimensions of the training data while maintaining high model accuracy. The optimization feature used is the TF-IDF (term frequency-inverse document frequency) feature extraction. sentiment analysis using TF-IDF feature extraction and stochastic gradient descent algorithm can classify Indonesian text appropriately according to positive and negative sentiment. Classification Performance using TF-IDF feature extraction and stochastic gradient descent algorithm obtained an accuracy is 85.141%.

Keywords: Covid-19, Twitter, Featured Extraction, Stochastic Gradient Descent.

1. Introduction

Twitter is a microblogging platform that can be used by every internet user and refers to the data of the Ministry of Communication and Information of the Republic of Indonesia [3] according to data

*Corresponding author

Email addresses: vindi.antonio@usu.ac.id (Vindi Dwi Antonio), syahril1@usu.ac.id (Syahril Efendi), mawengkang@usu.ac.id (Herman Mawengkang)

Received: May 2021 *Accepted:* October 2021

from PT Bakrie Telecom, that Indonesia has Twitter users with a total of 19.5 million out of a total of 500 million users in the world and continues to grow over time. The use of Twitter as a means of conveying information in the handling of covid-19 by the Indonesian government triggered Netizen to respond. Netizen's response consists of Positive and Negative.

Sentiment Analysis or opinion digging is the computational study of one's opinions, sentiments, emotions, judgments, and attitudes towards the entity of a product, service, organization, individual, problem, event, topic, and related attribute [8]. The technique for classifying sentiment analysis requires an algorithm. One of the classification algorithms is Stochastic Gradient Descent (SGD). According to Purwono [7], the Stochastic Gradient Descent Classification Algorithm is a simple and efficient approach in conducting linear classification with discriminatory learning. The SGD method is an iterative optimization algorithm for finding the minimum functionality points that can be derived. The algorithm starts by doing the tapping at the beginning of the process. The pushing error is then corrected as the guessing loop uses the gradient rule of the function to be minimized.

The more training data provided to the machine, the accuracy of the classification function model formed by the machine is also higher. But in making representations into numerical vectors, the dimensions of data become large due to the many features. Feature optimization needs to be done to the training data by reducing the dimensions of the training data while maintaining high model accuracy. The optimization feature used is the extraction of the TF-IDF (term frequency-inverse document frequency) feature. Extraction with TF-IDF (Term Frequency - Inverse Document Frequency) feature is one of the processes of feature extraction techniques with the process of assigning value to each word in the training data. To find out how important a word represents a sentence, it is calculated. The value of TF-IDF depends on the frequency of word occurrence in the document [5].

2. Related Work

In 2015 Chandani, Vinita e.t al researched Machine Learning Classification Algorithm Comparison and Feature Selection on Film Review Sentiment Analysis. From this research obtained the best algorithm comparison results are support vector machine algorithm with an accuracy of 81.10% and Area Under Curve of 0.904 [2].

In 2020 Amalia, Cindy, and Sibaroni, Yuliant conducted a study on the topic sentiment analysis on tweet data with Delta Weighting TF-IDF Using Artificial Neural Network Model. The results of this study stated that Delta TF-IDF weighting is better than the usual TF-IDF, as seen from the accuracy of all scenarios, Delta TFIDF obtained the highest accuracy results of 70.6% and TF-IDF of 68.5% [1].

3. Algorithm

We used the Stochastic gradient descent algorithm to classify sentiment analysis in this study. Stochastic Gradient Descent Classification Algorithm is a simple and efficient approach in conducting linear classification with discriminatory learning [7]. The classification process using SGD can be seen in Figure 1 below:

The SGD method is an iterative optimization algorithm for finding the minimum functionality points that can be derived. The algorithm starts by doing the tapping at the beginning of the process. The pushing error is then corrected as the guessing loop uses the gradient rule of the function to be minimized. The minimum function drop is used specifically with the formula below[6]:

$$\omega_i + 1 = \omega_i - \eta \nabla_{\omega_i} L(\omega_i)$$

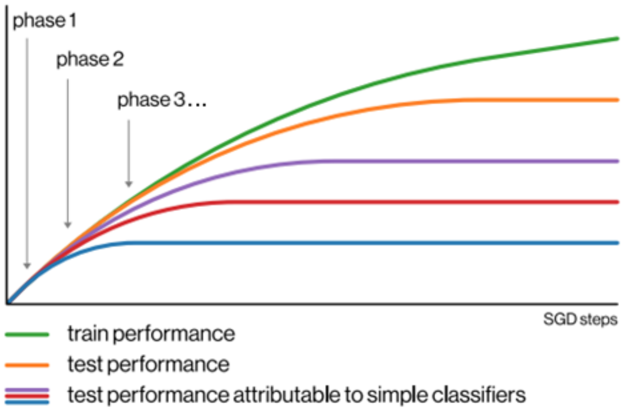

Figure 1: Stochastic Gradient Descent Classifier [4]

Description:

$\omega_i + 1$: predict parameter models
ω_i : parameter models on previous iteration
η : Learning rate
L : Loss cost function

SGD has a more label and faster nature when conducted classification training and is not limited to time in its implementation based on the size of the training dataset. The SGD method has faster learning abilities. The hinge loss function used as a classifier training can be explained by the formula below [6].

$$L(\mathcal{X}_j, \mathcal{Y}_j) = max(0, 1 - \mathcal{Y}_j \cdot (\omega\mathcal{X}_j + b)$$

Description:

ω *and* b : parameter models for predict
$\mathcal{X}_j$: input sample
$\mathcal{Y}_j$: target class

The equation above is a function of classification metrics as a measurement of linear model capabilities that have been predicted using the SGD method at each iteration of the learning phase. The equation is then modified parameters (w,b) resulting in a new equation. Classification using SGD, the value corresponds to the weight set for the reverse scatter feature of the decision function, and b is the intercept. An interesting part of the function of hinge loss is the process of punishing samples that are misclassified, but still given low trust as a barrier between classes.

Loss function also works with regularization that aims to help predicted models as well as generalize data that does not have labels. Regularization serves as a protocol for punishing complex models where more dominant overfitting occurs, which is characterized by a larger value for the Regularization parameter can be seen in the following formula

$$L1 = \sum_{i=1}^{m} |\omega_i|$$

$$L2 = \sum_{i=1}^{m} \omega_I^2$$

Description:

M : variable predictor
ω_i : parameter models to predict

4. Method

A typical workflow of our method can be described as:

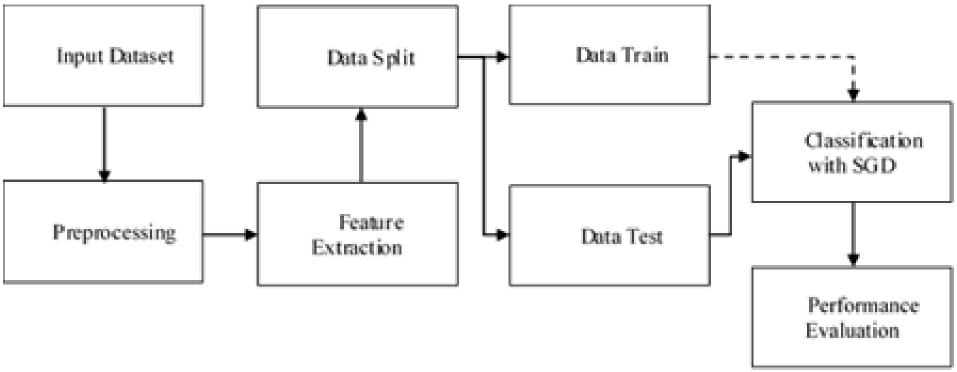

Figure 2: Workflow of our method

Figure 2 shows the overall workflow. Our study uses two basic programming tools. First, we use SQL to save datasets before and after preprocessing. Second, we use Python 3.9 to process the dataset until we got the sentiment analysis. use Tableau 8 to measure brand reputation and automate data visualization into the dashboard. Detailed processes are thoroughly explained in the following six points.

4.1. Dataset

In this paper, we use the dataset about covid-19 in Indonesia. The dataset to be used is an Indonesian Tweet from GitHub was crawling by Yahdi Indrawan, we can get it from his Github content URL: `https://raw.githubusercontent.com/yahdiindrawan/covid19-sentiment-dataset/master/Covid-19_Sentiment.xlsx`. The data used are only labeled positive and negative.

4.2. Preprocessing

In this research, the preprocessing stage uses the Sastrawi library in Python. This stage is done to avoid any data disruption during the weighting and classification process. The preprocessing used in this study includes cleansing, tokenizing, case-folding, stopword removal, and stemming. Detailed processes are thoroughly explained in the following points.

4.2.1. Cleansing

In this study, the cleansing process was conducted to remove tweet text that was considered unimportant. Here are the parts omitted from the tweet text like hashtags, mention, URL, retweet, and punctuation.

4.2.2. Tokenizing

At this stage, input strings will be cut based on each composed word.

4.2.3. Case-Folding

At this stage it is done to change all capital letters to lowercase letters, only the letter a-z is changed.

4.2.4. Stopword Removal

This stage is done to eliminate pre-words, pronouns, conjunctions, and words that have nothing to do with sentiment analysis.

4.2.5. Stemming

At this stage, the word healing will be converted into the base word.

After doing some of the above processes then the preprocessing stage is completed. Here's a table comparison of preprocessing results.

Table 1: Example of preprocessing comparison on tweet text

Tweet before Preprocessing	Tweet after Preprocessing
Pemprov Papua Naikkan Status Jadi Tanggap Darurat Covid-19. OPM	pemprov papua naikkan status jadi tanggap darurat covid 19 opm
Cegah covid-19 beserta jajaran Polsek Kuranji melakukan aksi peduli berupa pembagian masker gratis bagi pengguna jalan kegiatan ini dilakukan di depan Polsek Kuranji.Kamis (9/4)Febriputraguci	cegah covid 19 beserta jajaran polsek kuranji melakukan aksi peduli berupa pembagian masker gratis pengguna jalan kegiatan dilakukan depan polsek kuranji kamis 9 4febriputraguci

4.3. Featured Extraction

In the featured extraction process the tf-idf is done to convert the term into numerical numbers that will be processed as test data and training data. In this study, tf-idf feature extraction using sklearn module in python. The term conversion is done according to the number of text documents in the dataset. Here the X (input) is the text that has been preprocessed from the dataset and becomes Y (output) is a sentiment that is positive and negative.

4.4. Data Split

We split the dataset into 70% data train and 30% data test which is 189 tweets. We get 583 positive tweets and 246 negative tweets from all datasets which is 830 data tweets.

4.5. Classification

We classify the sentiment analysis with stochastic gradient descent.

4.6. Performance Evaluation

For performance evaluation, we got a confusion matrix to calculate accuracy, precision, and recall of stochastic gradient descent to classify the sentiment analysis. A confusion matrix is a table recording the results of classification performance. Table 2 is an example of a confusion matrix that classizes only two classes namely class 0 and 1. Each cell in the matrix represents the number of records/data from class i whose predicted results go to class j. For example, cell f: is the amount of data in class l that is correctly mapped to class l, and $f_{ij}$is the data in class 1 that is incorrectly mapped to class 0

Table 2: Confusion Matrix for 2 class prediction

f_{ij}		Predict Class (j)	
		Class = 1	Class = 0
True Class (i)	Class = 1	f_{11}	f_{10}
	Class = 0	f_{01}	f_{00}

5. Experiment

We used the Stochastic Gradient Descents (SGD) Algorithm as a classification algorithm for sentiment analysis of Indonesian tweets. the loss function used is the $L1$ function with the number of iterations as much as 10 times. In this study, the learning rate used was 0.5. We test 30 % of all datasets which is 249 datasets with 45 random states.

6. Result and Analysis

After conducting the test, the results of performance evaluation are obtained on each test presented in Table 3

Table 3: TF-IDF and SGD Feature Extraction Performance 10 iteration

Number of iteration	**Confusion Matrix (f_{ij})** f_{11}	f_{10}	f_{01}	f_{00}	**Accuracy %**	**Precision %**	***Recall %***
1	165	18	26	40	82,329	90,164	86,387
2	167	16	31	35	81,124	91,257	84,343
3	161	22	25	41	81,124	87,978	86,559
4	165	18	23	43	83,534	90,164	87,766
5	160	23	27	39	79,920	87,432	85,561
6	167	16	31	35	81,124	91,257	84,343
7	161	22	25	41	81,124	87,978	86,559
8	165	18	23	43	83,534	90,164	87,766
9	160	23	27	39	79,920	87,432	85,561
10	165	18	23	43	83,534	90,164	87,766
	Average				81,726	89,399	86,2611

7. Conclusion

Using featured extraction with term frequency and inverse document frequency and stochastic gradient descent algorithm can classify the sentiment analysis. And the best performance is with 4, 8, and 10 iterations on stochastic gradient descent algorithm with accuracy 83,534%. The result which we got is the performance of tf-idf featured extraction and stochastic gradient descent algorithm on classifying the sentiment analysis with accuracy is 81,726%.

References

[1] C. Amalia, and Y. Sibaroni, *Analisis sentimen data tweet menggunakan model Jaringan Saraf Tiruan dengan pembobotan delta Tf-idf.* eProceedings of Engineering, 7(2)(2020) 7810.

[2] V. Chandani, R. S. Wahono and P. Purwanto , *Komparasi algoritma klasifikasi machine learning dan feature selection pada analisis sentimen review Film.* Journal of Intelligent Systems, 1 (1)(2015).

[3] K. Kominfo, *Kominfo: Pengguna Internet di Indonesia 63 Juta Orang*, Kementrian Komunikasi dan Informatika, November, 7 (2013), `https://kominfo.go.id/`

[4] D Kalimeris, G Kaplun, P Nakkiran, B. Edelman, T. Yang, B. Barak and H. Zhang, *SGD on neural networks learns functions of increasing complexity,* Adv. Neural Inf. Process. Syst. , 32(2019) 3496-3506.

[5] A. M. Pravina, I. Cholissodin and P. P. Adikara (2019). *Analisis sentimen tentang opini maskapai Pepnerbangan pada Dokumen Twitter menggunakan algoritme support vector machine (SVM)*, Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 3(3)(2019) 2789-2797.

[6] A. S. Ritonga and E. S. Purwaningsih, *Penerapan Metode Support Vector Machine (SVM) Dalam Klasifikasi Kualitas Pengelasan Smaw (Shield Metal Arc Welding),* Edutic-Sci. J. Inf. Educ., 5 (1)(2018) 17- 25 .

[7] R. Umar, I. Riadi and Purwono, *Perbandingan metode SVM, RF dan SGD untuk penentuan model klasifikasi kinerja programmer pada aktivitas media sosial,* Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi) 4 (2)(2020) 329–335 .

[8] L. Zhang, S. Wang, B. Liu, *Deep learning for sentiment analysis: A survey*, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4) (2018) e1253 .