JISTE (Journal of Information System, Technology and Engineering)
Volume 1, No. 1, pp. 6-10
E-ISSN: -
http://gemapublisher.com/index.php/jiste
Received: 17-05-2023
Accepted: 17-05-2023
Published: 17-05-2023

The Application of Support Vector Machine Method to Analyze the Sentiments of Netizens on Social Media Regarding the Accessibility of Disabilities in Public Spaces

Handy Ferdiansyah [1], Nurul Komaria[2], Ilham Arief[3]

Universitas Muhammadiyah Sidenreng Rappang[1], Universitas Jember[2], STIKes Widya Dharma Husada[3]

Correspondence Email: handyferdiansyah888@gmail.com[1]

Abstract

According to the Ministry of Social Affairs of the Republic of Indonesia, based on data from the Central Bureau of Statistics, there will be 22.5 million people with disabilities in Indonesia. As time goes on, it is possible that the number of people with disabilities will increase. So, facilities and infrastructure that are friendly to people with disabilities must be given great attention. Various comments from the public regarding this matter arose on social media, especially Twitter, both positive and negative. There are lot of tweet in Indonesian related to this matter, which will be classified using the Support Vector Machine algorithm with the Radial Basis Function kernel using Grid Search and cross-validation. The parameters used are in the range C, γ and produce a maximum accuracy of the parameter combination, result of this study was aligned with previous research that a good and appropriate parameter combination of C and γ in the RBF kernel SVM method will produce maximum accuracy of the classification results.

Keywords: disabilities, infrastructure, social media, support vector machine.

INTRODUCTION

In their daily lives, people with impairments confront more difficulties than those without disabilities. This is because there are obstacles to the accessibility of public services, such as in terms of employment, transportation services, and others. Based on data from the Central Bureau of Statistics (BPS), people with disabilities in Indonesia until October 2020 reached 22.5 million people, or around 5% of Indonesia's population (Ministry of Social Affairs, RI, 2020). As time goes on, it is possible that the number of people with disabilities will increase. So, facilities and infrastructure that are friendly to people with disabilities must be given great attention. Disability should not be an obstacle to obtaining the right to life. The state must offer protection to regulate the status and rights of people with disabilities, according to Law Number 8 of 2016 respecting people with disabilities. However, in practice, in their daily lives, people with disabilities still face various problems. The lack of opportunities offered to them results in limited access to meet their needs (Borg & Boldt, 2020).

Various opinions about this emerged on social media. Social media is perceived to play a role in public deliberations. People use social media to express their opinions in response to this problem. One of the most popular social media platforms is Twitter. Social media users

in Indonesia, until February 2022, reached 58.3% of the total internet users aged 16 to 64 years, or nearly 112 million users (Hootsuite, 2022). One of its users is from the developer side. Social media is quite popular among developers because of its ease in retrieving the necessary data. In a post, it is limited to 140 characters so that users can convey their opinions in a short, concise, and clear manner (Andry et al., 2020). Opinions in the form of text can be collected and processed using a technique called sentiment analysis (Ceron & Memoli, 2016).

Sentiment analysis is a technique that examines people's perceptions of things including goods, services, issues, people, themes, events, and their characteristics. It also analyzes people's opinions, assessments, sentiments, attitudes, judgments, and emotions regarding those things (Tannady et al., 2020). Sentiment analysis itself is a data classification model with a supervised learning technique in machine learning that makes predictions that will be verified on test data events and their attributes using training data called training datasets (Gunawan et al., 2019). Sentiment analysis itself is a data classification model with a supervised learning approach in machine learning that uses training data known as training datasets to make predictions that will be tested on test data. In carrying out sentiment analysis, a qualified classification algorithm is needed in order to obtain maximum accuracy (Hadi, 2019). Support vector machine (SVM) for classification problems, linear regression for regression problems, and random forest for both classification and regression problems are some common supervised learning techniques (Hendy et al., 2020).

There are still numerous approaches to categorizing sentiment analysis. Support vector machine (SVM) is a prominent method that works well when employing a machine learning approach to sentiment analysis. due to the algorithm's popularity and the fact that few researches have made a comparison. Support Vector Machine (SVM), Multilayer Perceptron (MLP), Random Forest, Decision Tree, K-Star, and Naive Bayes were some of the techniques that underwent accuracy testing (Sultana et al., 2018). The results of this study show that the Support Vector Machine (SVM) method has the highest level of accuracy compared to these other methods in predicting student performance at 78.75%. In the process, the SVM method has a kernel trick where kernel functions are used to operate on a dimensional feature space that cannot be separated linearly (Suryono et al., 2018). In general, problems in the real world are rarely linear (Tannady et al., 2020). RBF (Radial Basis Function) kernels are typically a solid first option for model construction. In order to address nonlinear scenarios, the RBF kernel nonlinearly maps samples to a higher-dimensional space. The RBF kernel's process (C, γ) uses a number of parameters.

METHOD

At this research stage, the writer needs to collect data and information related to research to support the research process, as well as research reference material. The author searches for data and information online via the internet. The process of collecting data from related previous studies is carried out as follows: In a literature study, the author collects literature-related studies to serve as supporting data. The search for relevant data and information comes from books, previous similar research (research journals and theses related to the topics discussed by the author), and internet sites that can be used as references for this research. Researchers conduct field studies and data observations by utilizing social media to check the availability of data and social media API features for developers. The data is obtained by utilizing tokens and keys from the Social Media API as access keys to obtain social media data about public comments on public accessibility in public spaces that are still unfriendly or already friendly to people with disabilities in Indonesia. Data can be retrieved many times according to how much data is needed, but it can only be retrieved within a span of less than or the last 10 days.

Data crawling can be done using key codes and tokens obtained from the social media API and Google Collaboratory. Data is crawled using keywords related to the object under study and how much data is desired. If the data obtained does not meet the desired amount of data, then crawling the data can be done several times until the desired amount is met. The data obtained separately in each crawl will be put together in one file, which is then saved in CSV format. The file will then be pre-processed and converted into new data, which is then stored in a new file for pre-processing. Pre-processing of data is carried out to clean and homogenize existing data so that sentiment analysis is more optimal. After the Twitter data is preprocessed, the next process will be feature extraction. The dataset will be classified using the SVM approach after the data has been transformed into vector form. The LibSVM function, which may be used for SVM classification, will be used in the classification process. The Support Vector Classifier (SVC) with the Radial Basis Function (RBF) kernel will be the function that is used. By showing a classification report with the model's performance in predicting test data, including accuracy, precision, recall, and f1-score, the results of the SVM modeling will be assessed.

RESULT AND DISCUSSION

In the early stages, data is collected from a source called crawling data. The process of crawling data is carried out on social media. In order to crawl data on social media, users must have a social media account and a social media API. The social media API is used as a door for data retrieval in the form of uploads from social media. Furthermore, the dataset is processed in the preprocessing stage. Data preprocessing is a process or step that is important enough to clean data or reduce noise so that the data is more uniform and easier to process and analyze. This process consists of several stages, including: in the early stages, case folding is carried out to change uppercase letters to lowercase letters by utilizing the Pandas function. From this process, case folding results are obtained for each post. The following is an example of the results of case folding, which shows that uppercase or capital letters have been changed to lowercase. After the data has gone through the case folding stage, it will be cleaned of characters such as a mention, which mentions someone's account on social media; # hashtags, which are usually used as labels or topics in these posts; emoticons; double whitespaces; website codes; numbers; punctuation marks; and meaningless single characters using the Natural Language Tool Kit library in Python. The input used in this stage is the result of the case folding stage. From this process, a text is obtained that only contains alphabetic characters and single spaces. After that, the data will go through the tokenization stage, which is a process in which posts that were originally in the form of complete sentences will be separated into word-for-word or tokens based on the space characters in the post. The input in this stage is the result of the cleansing stage. From this process, the text is obtained by word

After the dataset has passed the tokenization stage, the data will be subjected to a stop word removal process, which is the process of removing words that are considered unimportant, meaningless, or have no effect on the classification process. This process uses the input from the previous tokenizing results and then matches it with the stop word dictionary as well as the additional stop words. If the token with the stop word list is the same, then the words or token will be deleted; if not, then the token is still the same as the initial form. For the stop word dictionary, the author uses an existing dictionary. The results of the stop word removal stage are tokens or words that contain important meanings, while meanings that are not important have been removed. The data that has been partitioned word by word is still in the original word form of the post, where the words are not necessarily standard words. So that at this stage, non-standard words will be converted into standard words by matching the words from the tokenization results with the words in the

normalization dictionary based on KBBI. If the token is found in the normalization dictionary, the token will be changed to the standard word according to what is contained in the dictionary. For the normalization dictionary, the user uses an existing dictionary and then adjusts it by adding a few words and their default form according to the available data.

Furthermore, the data is processed by changing the words to their basic word forms according to the KBBI standard. This stage uses the Nazief and Adriani algorithm (library literature). The input stemming process is the result of the normalization stage. The result of this stage is the root word of the previous word or a word that no longer has affixes. After the data is cleaned through the preprocessing stage, it will go through the labeling process. Labeling is done automatically using the Indonesian opinion dictionary Sentiment Lexicon, where each word has a polarity score from -5 to -1 for negative words and +1 to +5 for positive words. Words that are counted or processed are only words that are in the opinion dictionary; if the word on Twitter is not in the opinion dictionary, it will be skipped or ignored. If the calculation result from labeling is < 0 (less than 0), then it will be labeled negative; if ≥ 0 (more than equal to 0), it will be labeled positive. The results of the automatic labeling are then validated by manually checking as much as 10% of the existing data, or around 120 data points, and then reprocessing them so that the labeling results are more accurate. The following labeling results yield 776 positive data labeled '1' and 433 negative data labeled '0," for a total of 1209 data.

This stage uses the Scikit-learn Python library. The library has many modules, one of which is the TfidfVectorizer, which the author will use. After the Twitter data is labeled, the data will be given a term weight. This process gives value to each term in each post that has gone through the preprocessing stage. The value of the term will be input for the next stage, namely classification. Data that has been labeled so that it has a label and polarity becomes input for TF-IDF processing. In its application, the author sets max features to 2500 to get 2500 top terms with the largest term frequency. It is known that the greater the value or weight, the more often a term or word appears in the data. The word cloud results visualize words that frequently appear in the data. In the figure, it is known that the word 'disability' has the largest size, so it represents that these words appear most often in the data, so it has a large TF-IDF weight.

At the classification stage, the author tests the previous data that has been extracted by the feature extraction process. The model will later be tested to find out the level of accuracy of the model or the extent to which the model can classify test data. This process is called machine learning. Before the data enters the classification stage, it is initiated by X and y, with X being the Twitter data that has gone through the TF-IDF stage and y being the sentiment. The data will be classified by Stratified K-Fold Cross Validation as model validation with K = 10 because the higher the K, the higher the accuracy value. At this stage, the data will be split into 10 parts, with the first 10% of the subset as testing data and the remaining 90% as training data. This is done with iterations according to K, which is 10 times, with different test data and training data in each iteration. This stage also involves the distribution of training and test data. Each combination of parameters C and γ produced different accuracy, from the lowest 64.19% to the highest 83.37%. It is influenced by C and γ itself. C, as a parameter from the RBF kernel, is used to optimize the SVM method to avoid misclassification of each sample in the training dataset. If C is higher, then the decision boundary is smaller, so that the generalization properties of the classification (algorithm) are lost. Conversely, if C is lower, then the decision boundary is wider, so that it can be generalized properly but can classify some data as incorrect. While γ is how far the influence of one sample training dataset or training dataset. If γ is higher, then the decision boundary depends on the points closest to it. On the other hand, if γ is lower, then even distant points will be considered when deciding

where the decision boundary should be. Therefore, the combination of parameters C and γ must match the existing data. the highest accuracy of the combination of parameters C and γ with the accuracy results of each validation with Stratified K-Fold Cross Validation.

CONCLUSION

With positive and negative sentiment classification outcomes, the Support Vector Machine (SVM) method with the Radial Basis Function (RBF) kernel was successfully used to the sentiment analysis of Twitter about handicap accessibility in public settings. Tests in this study resulted in the maximum accuracy of classification using the SVM method using the RBF kernel for Twitter data about disability accessibility in public spaces obtained at parameters C = 1.000 and γ = 0.000001 (1e-06) of 83.37%. Develop keywords in posts if there are other issues related to disability accessibility in public spaces so that the data is more varied when studied. Adding sentiment to the dataset, for example, is neutral because, in this study, it only considers positive and negative sentiments. Consider POS tags such as negation and booster words in the calculation of labeling to avoid words that can cause ambiguity so that the labeling results are more accurate. Using parameter tuning on subsequent models to find the best combination of parameters for maximum accuracy results in.

REFERENCES

Borg, A., & Boldt, M. (2020). Using VADER sentiment and SVM for predicting customer response sentiment. *Expert Systems with Applications, 162*.

Andry, J. F., Tannady, H., & Gunawan, F. E. (2020). Purchase order information system using feature driven development methodology. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(2), 1107-1112.

Ceron, A., & Memoli, V. (2016). Flames and Debates: Do Social Media Affect Satisfaction with Democracy? *Social Indicators Research*, 126(1).

Gunawan, F. E., Andry, J. F., Tannady, H., & Meylovsky, R. (2019). Designing enterprise architecture using togaf framework in meteorological, climatological, and geophysical agency. *Journal of Theoretical and Applied Information Technology*, *97*(20), 2376-2385.

Hadi, I. (2019). The Urgency of Disability Accessibility in Gorontalo District Government Agencies. *Al-Himayah*, 3.

Hendy, T., Resdiansyah, R., Johanes, F. A., & Rustono, F. M. (2020). Exploring the role of ICT readiness and information sharing on supply chain performance in coronavirus disruptions. *Technol. Rep. Kansai Univ*, *62*, 2581-2588.

Sultana, J., Sultana, N., Yadav, K., & Alfayez, F. (2018). Prediction of Sentiment Analysis on Educational Data based on Deep Learning Approach. *21st Saudi Computer Society National Computer Conference, NCC 2018*.

Suryono, S., Utami, E., & Luthfi, E. T. (2018). Sentiment Classification on Twitter with Naive Bayes Classifier. *Angkasa: Jurnal Ilmiah Bidang Teknologi*, 10(1).

Tannady, H., Andry, J. F., Gunawan, F. E., & Mayseleste, J. (2020). Enterprise architecture artifacts enablers for it strategy and business alignment in forwarding services. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(2), 1465-1472.

Tannady, H., Andry, J. F., Sudarsono, B. G., & Krishartanto, Y. (2020). Enterprise architecture using Zachman framework at paint manufacturing company. *Technol. Reports Kansai Univ*, *62*(4), 1869-1883.