International Journal of Informatics Technology (INJIT)
Vol. 1., No. 2, May 2023
e-ISSN : 2985-9646
https://jurnal.amrillah.net/index.php/injit

Sentiment Analysis On Twitter Social Media Using Naïve Bayes Classifier With Swarm Particle Selection Feature Optimization And Term Frequency

Antoni Lukito[a,1,*], Hermawan Basuki[b,2]

[a]Universitas Muhammadiyah Surakarta

[b]Universitas Negeri Yogyakarta

[1]antony265@gmail.com *, [2]herm_86@gmail.com

* Corresponding Author

ABSTRACT

Today's social media users are very large, where everyone expresses opinions, comments, criticism and so on. This data provides valuable information that can help people or organizations in decision making. Very large amounts of data are impossible to divide humans to read and analyze manually. Sentiment Analysis is an internal process analyze, understand, and classify opinions, evaluations, judgments, attitudes, and emotions towards a particular entity such as a product, service, organization, individual, event, topic, in order to obtain information. This research aims to separate Indonesian language tweets on Twitter social media into positive, negative and neutral categories. Naïve Bayes Classifier (NBC) method with feature selection Particle Swarm Optimization (PSO) is applied to the dataset to reduce less relevant attributes during the classification process. The test results show that the Naïve Bayes Classifier algorithm with feature selection Particle Swarm Optimization (PSO) uses term frequency (TF) parameters with 97.48% accuracy.

INJITInternational Journal of Informatics Technology

KEYWORDS
Sentiment Analysis
Twitter
NBC
Term frequency
PSO

1. Introduction

The growth of online media such as media social Twitter encourages the emergence of information unlimited textual, so it appears the need to explore the value of information the. Textual information is categorized into fact and opinion. Facts are an objective expression regarding an entity, event or nature, while opinion is an expression subjective that describes sentiment a person, opinion or feeling about a entity, event or trait. On Twitter there is a term called a tweet which is a messages or statuses created by users. A tweet can express a feelings or circumstances of Twitter users. Tweets may contain an opinion from users regarding the events they experience.

This opinion can be used as assessment for a company or agency. Sentiment Analysis is part of Natural Language Processing (NLP) and Machine Learning. The way it works is by classify positive opinions, negative and neutral. Sentiment Analysis in analyze people's opinions, sentiments, evaluations, judgments, attitudes, and emotions toward an entity such as products, services, organizations, individuals, a problem, event or topic.

From several classification techniques most often used for data classification is the Naïve Bayes method that is often referred to with the Naïve Bayes Classifier (NBC). Algorithm Naïve Bayes was chosen because of this algorithm very suitable for short text data. Excess NBC is a simple method but has accuracy and high performance in text classification (Routray, et al, 2013). Feature selection is one of the factors most important thing that can influence the level classification accuracy due to if the dataset contains number of features, the dimensions of the space will be large, lowering the level of classification accuracy. Selection feature is an optimization process for reducing a large set of features of the original source order that a relatively small subset of features are significant to improve fast classification accuracy and effective.

Feature selection and settings parameters in NBC significantly affects the results of classification accuracy. So Therefore, this research uses combining feature selection and NBC methods. The feature selection used in this research is Particle Swarm Optimization (PSO) with term frequency (TF) parameters.

Sentiment Analysis has been carried out by several researchers. Sentiment analysis using Naïve machine learning techniques Bayes classifier with chi-square feature selection, to reduce internal disturbances (noise). classification. The research results show that expected frequency of feature occurrence in true category and in false category has an important role in feature selection chi-square. Then with classification obtained 83% accuracy and harmonic average 90.713% (Juen Ling, et al, 2014). His research combines methods SVM and PSO for classifying positive and negative opinions. Hybrid method (SVM - PSO) has improved accuracy, compared to using the method SVM only (K.Umamaheswari, Ph.D, et al, 2015).

Sentiment Analysis on Social Media Twitter Using Naive Bayes Classifier Regarding the Keyword "2013 Curriculum" (Dyarsa Singgih Pamungkas, et al, 2015). In This research uses the Twitter Search API retrieve data from Twitter, the author apply character n-gram process for selection features and uses the Naive Bayes algorithm Classifier for classifying sentiments automatic. The author uses 3300 tweet data about sentiment towards the keyword "curriculum 2013". The data is classified manually and divided into 1000 data each for positive, negative and neutral sentiment. For The training process uses 3000 tweet data and 1000 Tweet each sentiment category. Research result This results in a system that can classify sentiment automatically with test results of 3000 training data and 100 tweet data testing reached 91%. Twitter Sentiment Analysis for Text Speak Indonesian with Maximum Entropy and Support Vector Machine (Noviah Dwi Putranti & Edi Winarko, 2014). Analysis Sentiment in this research is a process classification of textual documents into two classes, namely positive and negative sentiment classes. Data opinions obtained from the social network Twitter based on queries in Indonesian.

This research aims to determine public sentiment towards a particular object delivered on Twitter in Indonesian, thus helping businesses to conduct research market over public opinion. The data is already there collected, preprocessing is carried out and POS tagger to generate models classification through the training process. Technique collection of words that have sentiment carried out using a based approach dictionary, which was produced in this research totaling 18,069 words. Maximum Algorithm Entropy is used for POS tagger and algorithm used to build classification models over deep training data This research is a Support Vector Machine.

The feature used is the unigram with TFIDF weighting feature. Implementation The classification accuracy obtained was 86.81% testing 7 fold cross validation for type Sigmoid kernel. Manual class labeling with POS tagger produces accuracy 81.67%.

2. Method

Feature selection and settings parameters in NBC significantly the results of classification accuracy. So Therefore, this research uses combining feature selection and NBC methods. Feature selection used in research this is Particle Swarm Optimization (PSO) with term frequency (TF) parameters. Text processing was carried out on the dataset first. Text preprocessing works to change text data that is not structured or arbitrary data structured. The process is carried out in stages preprocessing is as follows:

a. Case Folding

Case folding is an equalization process case in a document. This matter done to make things easier search. Not all text documents consistent in the use of capital letters. Hence the role of case folding needed for conversion the entire text in the document becomes a standard form (in this case letters small or lowercase).

b. Tokenizing

Tokenizing is a cutting process a document into parts, which is called a token. At the same time tokenizing too serves to dispose of some a particular character is considered to be punctuation.

c. Stopword Removal

Stopword removal is a process deletion of words that are not contributed much to the content of the document. The words included in stopword removed because have a bad influence in the text mining process such as words "how", "also", "so that", "so" and etc. The next stage is features selection. This stage is a stage important in text mining. One function The important things this process provides are: to be able to choose any term or word who can serve as an important representative for the collection of documents that we will analysis. In this research we will use feature selection, namely Particle Swarm Optimization (PSO) with term parameters frequency (TF). Particle Swarm Optimization (PSO) begins with a population consisting of a number of terms (which express solutions) that generated randomly and so on carry out a search for the optimal solution through term improvements for a number of certain categories.

Generation of position (xi, d) and velocity (vi, m) from a collection of particles generated randomly radom uses the lower limit (Xmin) and upper limit (Xmax). To update velocity (speed) for all terms of fitness value can be determined which terms have values global best (global best) and also can determine the best position (local best) for each terms at all times now and previously. Will do a repeat until the criteria are met. TF-IDF (Term Frequency – Inverse Document Frequency) is a numeric statistic which shows the importance of the word on document (Rajaraman, Leskovec, & Ullman, 2018).

Generally TF-IDF is used as factor to calculate the weight on Information retrieval. The TF-IDF value increases every time a word appears in a document, but it decreases if the word frequency is frequent appears, this is to handle the words that appears frequently. Therefore, the weight generated from TF-IDF can be used as one feature for performing Clustering/ grouping of words. TF can be demonstrated in 3 way (Salton & Buckley, 2020).

In this study, a novel MPONLP-TSA method is presented for the recognition of sentiments that exist in Twitter data during the COVID-19 pandemic. The presented MPONLP-TSA model performed data pre-processing to convert the data into a useful format. Following, the BERT model is used to derive word vectors. To detect and classify sentiments, the MPO algorithm with the BiRNN model is utilized. Figure 1 exemplifies the overall working process of the MPONLP-TSA method.

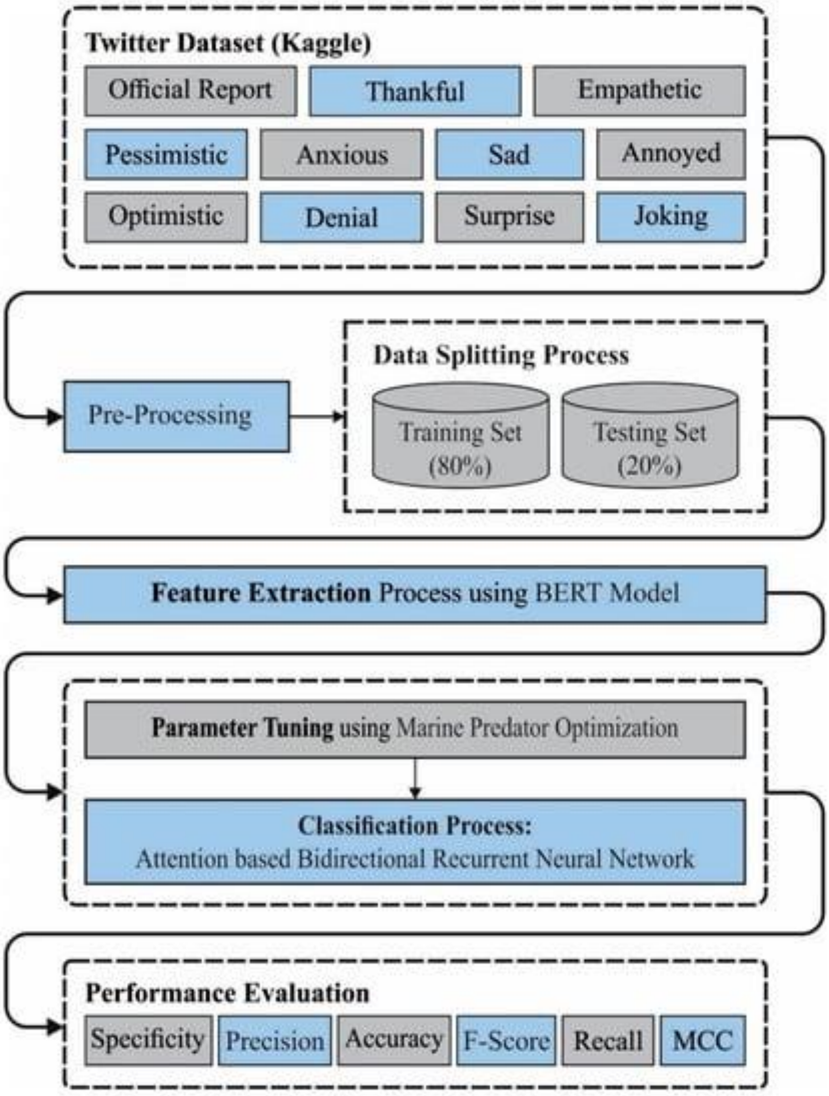

Fig. 1. Overall working process of the MPONLP-TSA method.

1. for the words contained in documents and a value of 0 against words which does not appear in the document. On this concept, the frequency of occurrence of words not included in the calculation.
2. Using the frequency of occurrence value said directly to become TF.
3. Use the fractional value of the term normalization has been carried out, the formula These are (Salton & Buckley, 1988): $TF(t, d)$ = 0.5 + 0.5 where (t,d) is the frequency of word t appears in documents d and max $\{f(w,d):w \in d\}$ is the frequency maximum of other terms in the document Meanwhile, to calculate the IDF The following formula is used (Salton & Buckley, 2021).

$$IDF(t, d) = \log$$

where N is the number of documents and $Df(t,D)$ as many documents in a collection of documents D containing the term t. However, if the term does not appear, it will There is a value of 0 in division, so it is necessary handling to replace it 1+$Df(t,D)$. To calculate the TF-IDF of a word, formula is used:

$$TF - IDF(t, d, D) = tf(t, d)x\ idf(t, D)$$

From the formula above you will get the value which can be used as word weighting when grouping words. Cross Validation stage, each record used several times for training and for testing. To illustrate this method, assume partitioning the data into two subsets. First, one of the two subsets is selected for training and another for testing. Then an exchange of functions is carried out subset such that the subset that previously as a training set into testing set and vice versa.

This method is a standard evaluation namely stratified 10-fold cross-validation because shows that 10-fold cross-validation is the best choice to get results accurate validation, 10-fold cross-validation will repeat the test 10 times and The measurement result is the average value of 10 testing times.

Classification Stage of Naive Bayes Algorithm The classifier is the algorithm used to find the highest probability value for classify test data into categories most appropriate (Feldman & Sanger, 2007).

In the naïve Bayes classifier algorithm each document is represented by attribute pair "x1, x2, x3,...xn" where x1 is the first word, x2 is the second word and so on. While V is a set Tweet category. During the classification algorithm will look for the highest probability of all category of documents tested (VMAP), where the equation is as follows:

For P(x1, x2, x3,...xn) the values are constant for all categories (Vj) so that the equation can be written as follows:

The above equation can be simplified be as follows Information :

Vj = Tweet category j =1, 2, 3,...n. Where in this study j1 = category negative sentiment tweet, j2 = positive sentiment tweet category, and j3 = tweet sentiment category neutral

P(xi|Vj) = Probability of xi in category Vj

P(Vj) = Probability of Vj

For P(Vj) and P(xi|Vj) it is calculated at during training where the equation is

3. Results and Discussion

In this research work, the number of hidden layers (1–3) and neurons (50–100), were assigned according to the number of features that become input for the MLP model. In iterations 1, 2, 3, and 4, a total of 50 neurons were taken into account for testing purposes using 1 hidden layer, and all the mentioned activation functions were gradually applied on this layer. This hidden layer was fully connected with one output layer to classify the sentiments as positive and negative labels. In contrast, from iterations 5, 6, 7, and 8, we have used two hidden layers on top of each other with 50,50 neurons as 50 neurons in each layer for testing purposes. Besides this, in iteration 9, 10, 11, and 12, only one hidden layer was used with 100 neurons for testing purpose, incorporating four activation functions one at a time. On the other hand, in iteration 13, 14, 15, and 16, we have used two hidden layers interconnected with each other, with 100,100 neurons as 100 neurons in each hidden layer for testing purpose. Added to this, in iteration 17, 18, 19, and 20, we have considered three hidden layers on top of each other with 50, 50, 50 neurons as 50 neurons in each hidden layer for testing purpose. In contrast, iteration 21, 22, 23, and 24 used three fully connected hidden layers, with 100,100,100 neurons as 100 neurons in each hidden layer, and four activation functions were applied for testing purpose to classify the sentiments that were taken as input from the feature selection method using input layer. Besides this, 4 activation functions were applied one at a time for each iteration and the number of neurons mentioned above. Depicting 4 activation functions in the same layer elucidates that all the activation functions were applied gradually.

Particle Swarm Optimization (PSO) used as feature selection and Naïve Bayes Classifier for evaluating feature subsets. In this study, we chose to use Particle Swarm Optimization (PSO) feature selection using the Term Frequency (TF) parameter with the Naïve Bayes Classifier because of the results feature selection obtained more terms, namely a total of 776 words were selected, meanwhile Particle Swarm Optimization (PSO) feature selection using TF-IDF parameters with Naïve Bayes Classifier obtained 774 terms selected words Number of feature selections with more terms many serve as important representatives for collection of documents to be analyzed. Particle feature selection test results Swarm Optimization (PSO) is proven to work improve the accuracy of the Naïve Bayes algorithm Classifier. Testing uses feature selection Particle Swarm Optimization (PSO) using term frequency (TF) parameters with Naïve Bayes Classifier of 97.48%.

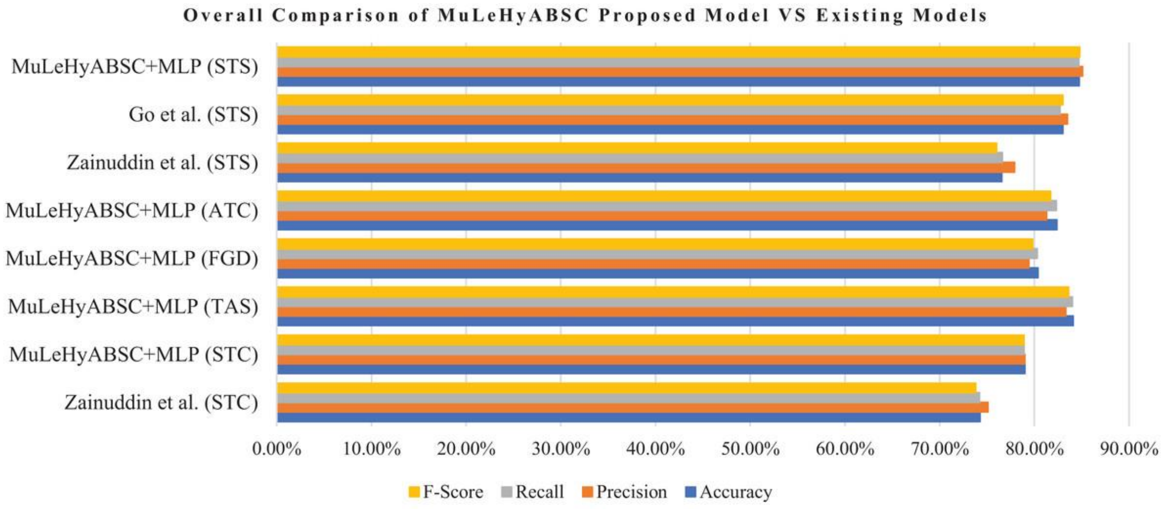

Fig. 2. System Accuracy Testing

This research work can be useful for other languages as well, like in the English language tweet datasets, we used POS tags as adjectives, verbs, and adverbs to find more implicit aspects and determined sentiment words describing them as shown in the proposed system MuLeHyABSC model in Fig. 1. In this work, the proposed approach performed better with the fusion of deep learning method MLP in all the datasets (used in this work), whilst machine learning approaches used for classification purposes didn't perform consistently, results varied in all the datasets. There are many possible reasons for inconsistency between results but one of the main reasons is variation in the sizes of datasets. In machine learning approaches, some classification methods performed better on small size datasets and some on large size datasets. In particular, POS tags +

unigram features and the combination of IG with PCA in feature ranking and feature selection process resulted in performance gain in the deep learning method.

The motivation of the proposed work was to perform finer-grained sentiment analysis to improve the functionality of aspect-based text classification using a hybrid approach. This study proposed an approach called: Multi-level Hybrid Aspect Based Sentiment Classification (MuLeHyABSC) comprising of multi-level (single and multi-word) aspect detection using ARM with the blend of heuristic POS patterns. The interconnection between noun phrases and heuristic mixture of POS patterns with verbs, adjectives, determiners, and adverbs was the main reason for valuable explicit aspects detection. Furthermore, the Stanford Dependency Parser (SDP) with the grammatical associations was used to find a relationship to extract implicit aspects including the determination of relations discovered by different types of dependencies. Our proposed approach also incorporates a feature ranking process by embedding a feature selection technique and further classification of sentiments using the deep learning method. This research aimed to use Twitter data to perform hybrid multi-level (single word and multi-word) aspect-based text classification. Different classification algorithms were implemented to compare the results with the proposed hybrid approach (MuLeHyABSC+MLP). The results showed that the proposed system for aspect-based text classification achieved significant improvement as compared to the existing baseline approaches proposed for sentiment classification by achieving accuracies of 78.99%, 84.09%, 80.38%, 82.37%, and 84.72% respectively. A neural network approach was used for large datasets which resulted in a performance gain. We plan to extend this research in the future by using some state-of-the-art approaches of ANN for aspect-based text classification. Latest feature extraction and feature selection techniques will be implemented with the combination of merged ANN methods including temporal aspects in the future for improving the existing system's performance.

4. Conclusion

Based on the results of the analysis carried out In research, things can be concluded the following. a. With feature selection in this research proven Particle feature selection method Swarm Optimization on datasets The total number of terms selected was 776 words, can help the Naïve classification process Bayes becomes more effective and accurate. The test results show that Naïve Bayes Classifier algorithm with feature selection using Particle Swarm Optimization (PSO) with term frequency (TF) parameters get 97.48% accuracy.

References

[1] Rahman, M.; Islam, M.N. Exploring the performance of ensemble machine learning classifiers for sentiment analysis of covid-19 tweets. In Sentimental Analysis and Deep Learning; Springer: Singapore, 2022; pp. 383–396. [Google Scholar]

[2] Chintalapudi, N.; Battineni, G.; Amenta, F. Sentimental analysis of COVID-19 tweets using deep learning models. Infect. Dis. Rep. 2021, 13, 32. [Google Scholar] [CrossRef] [PubMed]

[3] Mishra, R.K.; Urolagin, S.; Jothi, J.A.; Neogi, A.S.; Nawaz, N. Deep learning-based sentiment analysis and topic modeling on tourism during Covid-19 pandemic. Front. Comput. Sci. 2021, 3, 775368. [Google Scholar] [CrossRef]

[4] Costola, M.; Nofer, M.; Hinz, O.; Pelizzon, L. Machine Learning Sentiment Analysis, COVID-19 News and Stock Market Reactions; (No. 288), SAFE Working Paper; SAFE: Frankfurt am Main, Germany, 2020. [Google Scholar]

[5] Ebadi, A.; Xi, P.; Tremblay, S.; Spencer, B.; Pall, R.; Wong, A. Understanding the temporal evolution of COVID-19 research through machine learning and natural language processing. Scientometrics 2021, 126, 725–739. [Google Scholar] [CrossRef] [PubMed]

[6] Chandrasekaran, R.; Mehta, V.; Valkunde, T.; Moustakas, E. Topics, trends, and sentiments of tweets about the COVID-19 pandemic: Temporal infoveillance study. J. Med. Internet Res. 2020, 22, 1–12. [Google Scholar] [CrossRef] [PubMed]

[7] Khan, R.; Shrivastava, P.; Kapoor, A.; Tiwari, A.; Mittal, A. Social media analysis with AI: Sentiment analysis techniques for the analysis of twitter covid-19 data. Crit. Rev. 2020, 7, 2761–2774. [Google Scholar]

[8] Naseem, U.; Razzak, I.; Khushi, M.; Eklund, P.W.; Kim, J. COVIDSenti: A large-scale benchmark Twitter data set for COVID-19 sentiment analysis. IEEE Trans. Comput. Soc. Syst. 2021, 8, 1003–1015. [Google Scholar] [CrossRef]

[9] Alamoodi, A.H.; Zaidan, B.B.; Zaidan, A.A.; Albahri, O.S.; Mohammed, K.I.; Malik, R.Q.; Almahdi, E.M.; Chyad, M.A.; Tareq, Z.; Alaa, M. Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review. Expert Syst. Appl. 2021, 167, 114155. [Google Scholar] [CrossRef]

[10] Nemes, L.; Kiss, A. Social media sentiment analysis based on COVID-19. J. Inf. Telecommun. 2021, 5, 1–15. [Google Scholar] [CrossRef]