Jurnal Mantik, 6 (2) (2022) 1792-1798

Published by:Institute of Computer Science (IOCS)

Jurnal Mantik

Journal homepage: www.iocscience.org/ejournal/index.php/mantik/index

Classification Of Indonesian Slang Using Naïve Bayes And Decision Tree Methods On Social Media

Abdi Dharma[1], Aditya Calderon Naibaho[2], Lolo Mulatua Bancin[3], Alhoi Andrew Jefferson[4]

Program Studi Teknik Informatika, Universitas Prima Indonesia,Jl. Sekip Jl. Sikambing No.simpang, Sei Putih Tim. I, Kec. Medan Petisah, Kota Medan, Sumatera Utara 20111, Indonesia

Email : adityacldrn27@gmail.com

ARTICLE INFO

Article history:
Received: Jun 30, 2022
Revised: Jul 10, 2022
Accepted: Jul 18, 2022

Keywords:
Classification,
Naïve Bayes,
Decision Tre,
Social Media

ABSTRACT

At this time the application of language that appears on social media is the application of slang as a character of current language development. This phenomenon deserves to be discussed because it can find out how much slang is used on social media. The source of the dataset for this research is the verbal form found on the author's personal social media such as Instagram and Tiktok obtained by the web scraping method as many as 2,000 samples and the data will be divided into two categories, namely the category of slang and non-slang. This study aims to compare two classification algorithms, namely Naïve Bayes and Decision Tree to see which algorithm is more effective in classifying how many social media users use slang in commenting based on the dataset we have collected, so that results are obtained to see how high the percentage of usage is. Indonesian people's slang in commenting on social media.

1. Introduction

Language is a very important part of human life because humans are used to communicating with each other easily through language. Surprisingly, the content of language depends on both the words one uses and the way those words are expressed. If a speaker can't make his or her point in one language, they must switch to a more comprehensible one. Currently, slang has become a phenomenon in the era of the millennial generation, not only for people who live in cities as well as those who live in villages who know slang, language seems to have become a trend among millennials. On the other side, parents worry how challenging it is to grasp their children's speech and language. Additionally, social media is expanding right now, moving beyond Short Message Service (SMS) to include Instagram, Facebook, WhatsApp, Twitter, and other platforms [1].

Slang, according to Hornby (1974), are words, phrases, and word meanings that are frequently used in speech among friends or coworkers but are inappropriate for professional settings or excellent writing. Slang is the language (words, phrases, and uses) of informal registers that members of a special group such as youth, musicians, or criminals support to establish group identity. Slang is usually popular among millennials [2].

The Naïve Bayes algorithm is an algorithm for classifying data based on the Bayes theorem assuming the independence of the parameters from each other. By considering the value of another parameter, Bayes' theorem offers a method for estimating the likelihood of a parameter's value. [3]. Probability based on exclusive features in the data arises as a member of the probability sequence and is obtained by calculating the frequency of each class feature value based on the construction of the dataset. The training dataset is a

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

subset that is used to train classification solving procedures. The training process uses known values to predict unknown values [4].

Decision Tree algorithm is a classification method that is generally used in various types of fields, such as machine learning, image processing, and pattern identification [5]. Because it is straightforward and simple to grasp, this algorithm has been the most often applied and well-liked model. The decision tree is generated from the training data in the top down direction. The starting phase of the first decision node tree is the root node of decision tree. The tree's nodes each hold information. On the basis of an algorithm, several calculations are completed and the decision tree nodes are divided into two or more branches. In some cases, the node cannot be split, in which case it will be the final decision node [6].

2. Methods

Classification is a process for designing models used to group labels. Using unlabeled data, the classification process aims to produce label predictions. The classification process designs a model that can classify data to get a certain accuracy and level of precision with a label as a supervisor in the training process.

The Naive Bayes method and the Decision Tree method are employed for classification. However, before doing the classification, there are steps that must be done to complete research on the classification system, namely data collection, data preprocessing, implementation of the Naive Bayes algorithm and the Decision Tree algorithm.

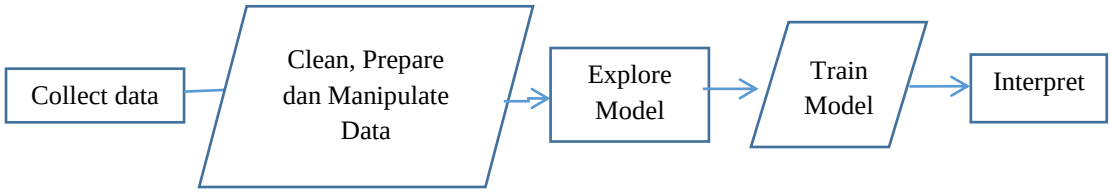

Figure 1. Data Science Flowchart

2.1 Dataset Collection

Researchers use a data collection technique, one of which is web scraping. Web scraping is widely applied to obtain datasets in the form of text from websites. This technique can facilitate researchers in the process of collecting datasets. The dataset used in this study is comments from social media Instagram and Tiktok in the period August 2021 – December 2021 as many as 2,000 data records.

TABLE 1
Dataset

No.	Name	Date	Likes	Comment	Target
1	wahyum41	20/09/21 06:50:46	0	bengek 😂😂😂😂	YES
2	Tyanangel	20/09/21 23:58:07	188	gua malah nerusin Lagunya:)	YES
3	rizki_jawsky	21/09/21 00:45:38	9	gw kok jadi lupa lirik aslinya	YES
4	Nurulzyi	21/09/21 00:46:04	45	yng terakhir kayak lagu kaset motocros adek gue dulu 😂	YES
5	deagiri0	21/09/21 01:21:26	106	bngekk bett smpe skt prt gwa ktawa😂😂😂	YES
…	…	….	…	…	…
2000	veronicha_meirin	12/18/2021 8:09:07	61	Perlu ada ini di Indonesia, luar biasa.	NO

In this study, the data to be processed in the dataset is the Comment column, we manually label to separate Comments that contain slang and those that don't, Targets with YES labels for Comments that use slang vocabulary, and Target NO for Comments that use vocabulary raw.

2.2 Preprocessing Data

The data preprocessing process is used so that the dataset is clean from noise, has a smaller size, is structured, so that further processing can be carried out. The preprocessing stage consists of Case Folding, Tokenizing, and Stopwords Removal.

2.3 Naïve Bayes Algorithm

Naïve Bayes algorithm is a classification algorithm used to predict the probability of each class [7]. This method takes advantage of the theory put forward by British scientist Thomas Bayes 8, which predicts future

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

probabilities based on past phenomena. Bayes' theorem is a useful formula in statistics for calculating the probability of a prediction. This algorithm calculates the probability of a class from each existing group attribute, and determines the most optimal type of class [8].

The Naive Bayes algorithm is a straightforward yet highly accurate supervised learning technique that can classify data.

$$P(C_i|X) = \frac{P(X|C_i).P(C_i)}{P(X)}$$

explanation :
X : data with unknown class
C_i : data hypothesis X is a specific class
$P(C_i|X)$: hypothesis probability C_i based on condition X (posteriori probability)
$P(C_i)$: hypothesis probability C_i (prior probability)

2.4 Decision Tree Algorithm

A classification technique called a decision tree has a tree-like form, with the leaves serving as classes and each node representing an attribute and a branch representing an attribute value [9]. The node at the very top of this decision tree is known as the root. A decision tree is a tree in which each node (or attribute) is a feature, each link (or branch) is a decision (or rule), and each leaf is a consequence (categorical or continuous value). Because decision trees mimic human level thinking it is very easy to take data and make it good. For all data, a tree of this kind should be created, with each leaf processing one result [10].

3. Results and Discussion

3.1 Result of Data Pre-processing

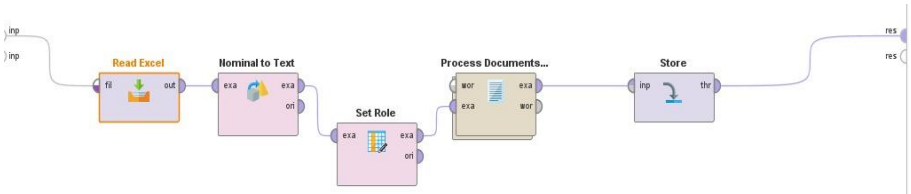

Figure 2. Text Pre-processing

In this pre-processing stage, selection of datasets that are not constant, duplicate data, and incomplete (missing value) will be carried out. This process is done by selecting non-constant data and deleting incomplete data. The dataset owned by the researcher is 2000 data. The dataset will be used for data sharing as training and test data. Then the data will be processed using rapidminer studio version 9.10 software which can generate a prediction.

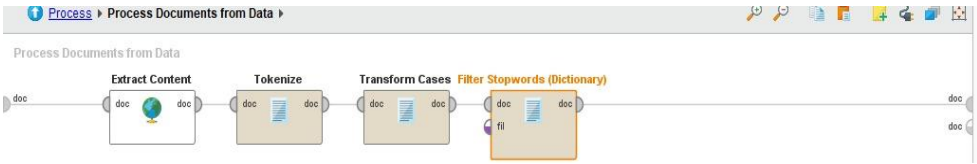

Figure 3. Process Documents from Data

TF-IDF vector main creation is used during the data pre-processing stage. Each word in the sentence is given a weight via TF-IDF, which then determines its inverse value. Each element of the document is described in words. The TF-IDF is represented as an array, where the columns of the array are words or features and the rows of the array are data. The obtained weight serves as a classification input [11]. The stages of pre-processing data include the following stages :

a. Tokenization : breaking up text into tokens or specific parts in the form of phrases, paragraphs, or documents.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

b. Transform case : This process is used to convert the token/text from the tokenization process to lowercase.
c. Filter Stopwords : This process is needed to remove or filter text that is not important in a sentence.

3.2 Algorithm Test Results

This test is carried out in the model development process using training data and validation data. After that, the classification process is carried out with the model and data testing to obtain the values of accuracy, precision, and recall based on the performance vector.

3.3 Naïve Bayes Algorithm Test Results

In the process of testing this model, the nave Bayes algorithm is used to perform the classification process on our dataset. Figure 4. is the flow of the classification process using the nave Bayes algorithm using rapidminer software.

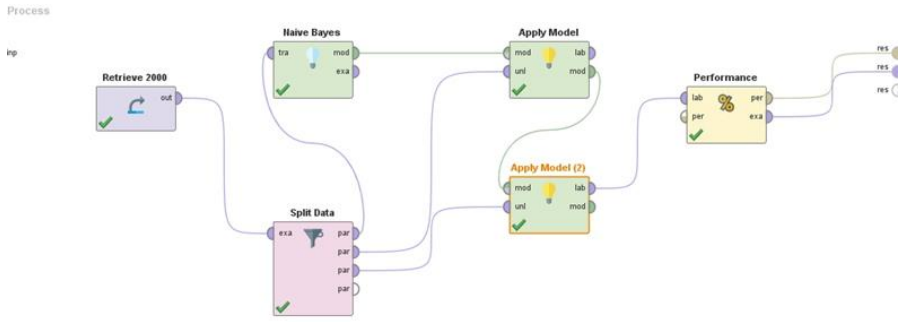

Figure 4. Flow Classification Naïve Bayes

The data used in the classification of this model consists of 2,000 datasets, 60% of which is used as training data and 20% is used for validation data and 20% is used as testing data. We did the test using and got the following result :

accuracy: 67.75%

	true YES	true NO	class precision
pred. YES	203	31	86.75%
pred. NO	98	68	40.96%
class recall	67.44%	68.69%	

Figure 5. Confusion Matriks in Naïve Bayes

The percentage of accuracy obtained by using the Naïve Bayes algorithm is 67.75%. The percentage of class recall for YES results is 67.44%, while the percentage of class recall for NO results is 68.69%. The percentage of class precision for pred.YES is 86.75%, while the percentage of class precision for pred.NO is 40.96%. To test this value, the calculation process can also be done manually, which is as follows :

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)}$$
$$= \frac{(203 + 68)}{(203 + 98 + 31 + 68)}$$
$$= \frac{271}{400}$$
$$= 0.6775 = 66.75\%$$
$$Precision = \frac{(TP)}{(TP + FP)}$$
$$= \frac{(203)}{(203 + 98)}$$
$$= \frac{203}{301}$$
$$= 0.6744 = 67.44\%$$

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

$$Recall = \frac{(TP)}{(TP + FN)}$$
$$= \frac{(203)}{(203 + 31)}$$
$$= \frac{203}{234}$$
$$= 0.8675 = 86.75\%$$
$$F1\ Score = \frac{(2 * Recall * Precision)}{(Recall + Precision)}$$
$$= \frac{(2 * 0.8675 * 0.6744)}{(0.8675 + 0.6744)}$$
$$= \frac{1.170084}{1.5459}$$
$$= 0.7588 = 75.88\%$$

3.4 Decision Tree Algorithm Test Results

In the process of testing this model, the Decision Tree algorithm is used to carry out the classification process on our dataset. Figure 6. is a flow classification process using a decision tree algorithm using rapidminer software. In the decision tree model, this study uses the Gini index criterion.

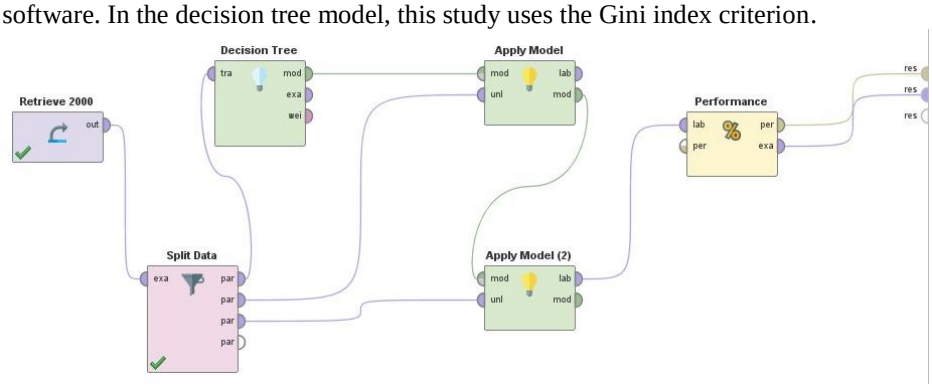

Figure 6. Flow Classification Decision Tree

The data used in the classification of this model consists of 2,000 datasets, 60% of which will be used as training data and 20% used for validation data and another 20% as testing data. We did the test using and got the following results :

accuracy: 77.25%

	true YES	true NO	class precision
pred. YES	291	81	78.23%
pred. NO	10	18	64.29%
class recall	96.68%	18.18%	

Figure 7. Confusion Matriks in Decision Tree

The percentage of accuracy obtained using the Decision Tree algorithm is 77.25%. The percentage of class recall for YES results is 96.68%, while the percentage of class recall for NO results is 18.18%. The percentage of class precision for pred.YES is 78.23%, while the percentage of class precision for pred.NO is 64.29%. To test this value, the calculation process can also be done manually, which is as follows :

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)}$$
$$= \frac{(291 + 18)}{(291 + 10 + 81 + 18)}$$

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

$$= \frac{309}{400}$$
$$= 0.7725 = 77.25\%$$

$$Precision = \frac{(TP)}{(TP + FP)}$$
$$= \frac{(291)}{(291 + 10)}$$
$$= \frac{291}{301}$$
$$= 0.9668 = 96.68\%$$

$$Recall = \frac{(TP)}{(TP + FN)}$$
$$= \frac{(291)}{(291 + 81)}$$
$$= \frac{291}{372}$$
$$= 0.7823 = 78.23\%$$

$$F1\ Score = \frac{(2 * Recall * Precision)}{(Recall + Precision)}$$
$$= \frac{(2 * 0.7823 * 0.9668)}{(0.7823 + 0.9668)}$$
$$= \frac{1.51265528}{1.7491}$$
$$= 0.8648 = 86.48\%$$

3.5 Discussion

This study compares the decision tree and the Naive Bayes classification model. In this study, the decision tree algorithm got better accuracy than the nave Bayes algorithm. A flaw in the Naive Bayes method is the choice of qualities that can influence the accuracy value [4]. While there may be certain issues with the Decision Tree's resilience, flexibility, scalability, and height optimization. However, Decision Trees produce a set of effective rules that are simple to comprehend [5]. The results of this study indicate that Indonesian people use slang more than standard language on social media to communicate.

4. Conclusion

Based on the findings of the research, it is possible to conclude : When using the Decision Tree algorithm, we get an accuracy of 77.25%. With the class recall value is 78.23%, the class precision value for is 96.68%, and the F1 Score value is 86.48%. When using the Naïve Bayes algorithm we get an accuracy of 66.75%. With the class recall value is 86.75%, the class precision value for is 67.44%, and the F1 Score value is 75.88%. Based on the results obtained, it can be ascertained that the classification model of slang language data management using the Naïve Bayes algorithm and the Decision Tree algorithm using the Decision Tree algorithm is more accurate in classifying data than the Naïve Bayes algorithm.

References

[1] Rezeki TI, Sagala RW. Semantics Analysis of Slang (SAOS) in Social Media of Millennial Generation. KREDO J Ilm Bhs dan Sastra. 2019;3(1).
[2] Rezeki TI, Sagala RW. Slang Words Used By Millennial Generation in Instagram. J Serunai Bhs Ingg. 2019;11(2):74–81.
[3] Handayani I, Ikrimach I. Accuracy Analysis of K-Nearest Neighbor and Naïve Bayes Algorithm in the Diagnosis of Breast Cancer. J Infotel. 2020;12(4):151–9.
[4] Wibawa AP, Kurniawan AC, Murti DMP, Adiperkasa RP, Putra SM, Kurniawan SA, et al. Naïve Bayes Classifier for Journal Quartile Classification. Int J Recent Contrib from Eng Sci IT. 2019;7(2):91.
[5] Charbuty B, Abdulazeez A. Classification Based on Decision Tree Algorithm for Machine Learning. J

Appl Sci Technol Trends. 2021;2(01):20–8.
[6] Singh M, Vyas M, Chaudhary R, Soni US. DECISION TREE ACADEMIC PERFORMANCE MODEL. J Posit Sch Psychol. 2022;6(2):4903–7.
[7] Hasudungan R, Pranoto WJ, Rudiman. Using MDA to Improve Naïve Bayes Classification for Students Performance Prediction. JSE J Sci Eng. 2020;1(2):65–70.
[8] Normah N. Naïve Bayes Algorithm For Sentiment Analysis Windows Phone Store Application Reviews. SinkrOn. 2019;3(2):13.
[9] Sulistiani H, Aldino AA. Decision Tree C4.5 Algorithm for Tuition Aid Grant Program Classification (Case Study: Department of Information System, Universitas Teknokrat Indonesia). Edutic - Sci J Informatics Educ. 2020;7(1):40–50.
[10] Patel HH, Prajapati P. Study and Analysis of Decision Tree Based Classification Algorithms. Int J Comput Sci Eng. 2018;6(10):74–8.
[11] Luthfi MF, Lhaksamana KM. Implementation of TF-IDF Method and Support Vector Machine Algorithm for Job Applicants Text Classification. J Media Inform Budidarma. 2020;4:1181–6.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.