SENTIMENT ANALYSIS WITH A CASE STUDY OF PRACTICE CARD ON TWITTER SOCIAL MEDIA USING NAIVE BAYES METHOD

Asthakhuroh[1*], Rachman Komarudin[2], Desiana Nur Kholifah[3]

Sistem Informasi[1,2,3]
Universitas Nusa Mandiri[1,2,3]
www.nusamadiri.ac.id
asthakhuroh1@gmail.com[1*], rachman.rck@nusamandiri.ac.id[2], desiana.dfh@nusamandiri.ac.id[3]

Abstract*—In early March 2022, Indonesia experienced a coronavirus pandemic which caused COVID-19 to enter for the first time. Since then, all sectors have been affected by the COVID-19 pandemic, not only health, the economic sector has also been seriously affected by this pandemic. In overcoming employment problems, the government makes a policy of the Pre-Employment Card program. The Pre-Employment Card Program is one of the government's efforts to expand job opportunities and to increase competitiveness which later became one of the social assistance for the community to overcome the Covid-19 pandemic. The implementation of the Pre-Employment Card program received pros and cons from the community, one of which was on Twitter social media. The results of the sentiment analysis of the pre-employment card program are mostly positive. The test results show that the Naïve Bayes Classifier method is successful in classifying sentiment with the highest accuracy value of 96%, the highest precision value of 98%, and the highest recall value of 96%, and AUC of 96%.*

Keywords*: Twitter, Naive Bayes Classifier, Sentiment Analysis, precision, recall*

Intisari— Pada awal Maret 2022 Indonesia mengalami pandemi virus corona yang menjadi penyebab COVID-19 masuk untuk pertama kalinya. Sejak itu seluruh sektor terdampak dari pandemi COVID-19 tak hanya kesehatan, sektor ekonomi juga mengalami dampak serius akibat pandemi ini. Dalam mengatasi masalah ketenagakerjaan pemerintah membuat kebijakan program Kartu Prakerja. Program Kartu Pra Kerja adalah salah satu bentuk upaya pemerintah dalam memperluas kesempatan kerja dan untuk peningkatan daya saing yang kemudian menjadi salah satu bantuan sosial bagi masyarakat guna penanggulangan pandemi Covid-19. Pelaksanaan program Kartu Prakerja mendapat pro dan kontra dari masyarakat salah satunya pada media sosial twitter. Hasil analisis sentimen program kartu prakerja kebanyakan bersifat positif. Hasil pengujian menunjukkan bahwa metode Naïve Bayes Classifier berhasil mengklasifikasi sentimen dengan nilai akurasi tertinggi sebesar 96%, nilai precision tertinggi sebesar 98%, dan nilai recall tertinggi sebesar 96%, serta AUC 96%.

INTRODUCTION

The government is trying to solve problems related to the number of employees who have been laid off, and the increasing unemployment rate which is now being discussed by the public. Participants of the Pre-Employment Card Program receive training fees from the government to improve their abilities so that they can be absorbed by the appropriate organization or create their own line of business as an example of efforts in this direction (Perekonomian, 2020). This program is designed to help people who are looking for work, as well as those who are currently employed but need to upgrade their skills because they have been laid off.

Targeting a workforce with the qualities and skills required to compete in the Indonesian labor market is the main objective of this initiative by the government. The Pre-Employment Card focuses on skills training needed in the current industrial era so that it can match the desired needs and be able to produce a workforce that is able to compete.

Although it has a constructive purpose, in some cases the Pre-Employment Card is considered a new problem by the general public. People use Twitter as a means of communication and gathering information every day. Twitter is used for various information or tweets between users so that they can share information in real time.(Fikri, Sabrila, & Azhar, 2020)

The existence of Twitter social media makes netizens share information about the latest topics being discussed, for example about Pre-Employment Cards starting from the registration process to the implementation of training. Sentiment analysis on Twitter regarding the Pre-Employment Card is very important to analyze the opinions of netizens regarding the Pre-Employment Card in the form of tweets, re-tweets and existing comments. Tweet data about the Pre-Employment Card on Twitter will later be analyzed to find out positive and negative comments so that information about the sentiments of Twitter netizens regarding the Pre-Employment Card can be generated.

More or less inspired and references from previous research related to the background of the subject in this thesis were referred to during the making of this thesis. The following studies, among others, have relevance to this research:

Rachmawan Adi Laksono et al 2019 carried out research entitled "Sentiment Analysis of Restaurant CustomerReviews on TripAdvisor using Naive Bayes". This study tries to classify customer satisfaction in Surabaya restaurants using Naive Bayes. A sampling of crawled data using WebHarvy Tools. The results of this study indicate that both methods obtain accurate customer responses and the Naive Bayes method is more accurate than TextBlob sentiment analysis with an accuracy difference of 2.9% (Laksono, Sungkono, Sarno, & Wahyuni, 2019).

Pandhu & Diki in 2020 conducted a study entitled "Sentiment Analysis and Classification of Positive Comments on Twitter with Naive Bayes Classification" which discussed the analysis of positive and negative English sentiments classified using the Naive Bayes Classification which was downloaded from the Sentiment 140 site. (Pandhu & Diki, 2020)

In 2019, Ruhyana conducted a study entitled "Analysis of Sentiment Against the Application of Odd/Even Number Plate Systems on Twitter Using the Naive Bayes Classification Method." This study explores the application of odd/even techniques to sentiment analysis on Twitter to categorize public sentiment on Twitter social media. This research is modeled with Rapid Miner Studio and preprocessed the data by deleting URLs, replacing emoticons and negation. This study uses data mining methods for classification with the Naive Bayes Classifier algorithm to classify Twitter users into positive and negative attitudes towards odd-even policies. The Naive Bayes Classifier (NBC) is a method based on Bayesian probability to solve data grouping.(Ruhyana, 2019)

This research conducted by Rosdiana in 2019 with the title "Analysis of Sentiment on Twitter towards Makassar City Government Services" discusses sentiment analysis regarding government services in Makassar City based on tweet data contained on Twitter. This study uses the Python programming language for data retrieval with the help of the tweepy library which produces json-shaped tweet data and then stored in a software called Elasticsearch. In this study, the k-fold cross validation test is divided into 3, 5 and 10 k subsets to obtain the desired accuracy results. This method uses Naive Bayes as a sentiment classification method and the output is sentiment with positive, negative, and neutral sentiments.(Rosdiana, Eddy, Zawiyah, & Muhammad, 2019)

Buntoro in 2017 carried out a research entitled "Analysis of Sentiments for the 2017 DKI Jakarta Governor Candidates on Twitter" discussing the analysis of sentiments of Twitter users related to the 2017 DKI Jakarta Pilkada with the keywords AHY, Ahok and Anies. This study performs several stages of analysis such as data preprocessing, tokenization, Part of Speech (POS) Tagger and classification using Naïve Bayes Classifier (NBC) and Support Vector Machine.(Buntoro, 2017)

In Rosit Sanusi et al's research, the data used was crawled using the help of the Twitter API which was taken from April 2020 to January 2021 as many as 4122 tweets. The research resulted in a system capable of classifying sentiments (positive, neutral, and negative) on a tweet. The accuracy level of the testing process is 64.48%. Some of the obstacles in the sentiment analysis process are the data making the model unbalanced, causing overfitting (Sanusi, Astuti, & Buryadi, 2021).

This study tried to use the Naive Bayes method which has a lower error rate when the dataset is large, besides that the accuracy and speed are higher when applied to a larger dataset using a balanced dataset.

Based on the background of the problems above, a research related to Sentiment Analysis With A Case Study Of Practice Card On Twitter Social Media Using Naive Bayes Method. Where it is necessary to do a research on sentiment analysis regarding the Pre-Employment Card on Twitter using the naive Bayes method to find out the positive and negative views of netizens regarding the Pre-Employment Card so that it helps in determining information about the Pre-Employment Card program so that it can be a separate consideration in the implementation of services.

MATERIALS AND METHODS

I. Literary Studies

a. Text Mining

Text mining is an approach or can be said to be a computer-based algorithmic technique, while the main function is to obtain new knowledge hidden from a set of texts. Text mining can also be said to be part of scientific information retrieval that works on data. with text types that tend or are unstructured (Priyanto, 2018).

b. Sentiment Analysis

As a business intelligence tool, sentiment analysis has been used. Unstructured text containing public opinion on a particular topic, product or service can be extracted using sentiment analysis. Sentiment analysis is a method to understand, extract and process textual input automatically to obtain reliable information(Salim & Mayary, 2020)

c. Naive Bayes

Naïve Bayes Classifier is a classification technique based on the Bayes theorem discovered by British scientist Thomas Bayes. This theorem works based on the independent assumptions of predictors so that future opportunities can be identified based on previous experience. In the Naïve Bayes algorithm, a number of clues called attributes are needed to assist in forming the appropriate class for the sample being analyzed (Artha et al., 2018; Watrianthos, Suryadi, Irmayani, Nasution, & Simanjorang, 2019)

In determining the value of the Naïve Bayes equation, formula (1) can be used as follows:

$$(C|F1,\ldots,Fn) = \frac{p(C)p(F1,\ldots,Fn|C)}{P(F1,\ldots Fn)}$$

After feature formation, the process is continued by calculating the probability of each class, which is done in equation (2):

$$\text{p (ci)} = \frac{fd\ (Ci)}{|D|}$$

Description of equation (2):
fd (ci) = the number of documents in the class
ci | D = the amount of training data

When the probability of each class has been obtained, then the probability calculation of each feature in the sentiment class is done with equation (3) below:

$$\text{p (wk|ci)} = \frac{f(wki, ci) + i}{f(ci) + |W|}$$

Description of equation (3):
f (wk, ci) = value of occurrence of the word wk in ci class
f (ci) = total number of words that appear in ci class
|W| = total number of wk

d. Evaluation Model & Classifier

The Confusion Matrix is a tabular representation of the predictive performance of a supervised learning algorithm. In the confusion matrix table, the data for each class shows the number of classification predictions made to classify the class as true or false . The confusion matrix has 4 terms as representations in classification or data modeling including "True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN)." Negative data that is correctly identified as negative is called True Negative (TN), while data that is incorrectly identified as positive is called False Positive (FP).(Khalimi, 2020)

e. Twitter

Indonesian people often express themselves by using social media as a means of communication. One of the media that is often used by this group is Twitter. Twitter is a social media that provides a lot of information through tweets, from the information written there is data that can be processed. (Ananda & Pristyanto, 2021)

II. Research methods

The research method used is the following data sources:

a. Literature Study

Data collection method using the Naive Bayes Classification approach uses library sources such as journal articles and literature, books, and internet sites as library sources for writing materials.

b. Observation

By using the Naive Bayes Classification approach, both primary and secondary data were examined qualitatively.

c. Design

The design of the data retrieval flow and the measurement of the data flow are the two main aspects of this approach.

d. Training and Classification

At this stage, data training is carried out using the Naive Bayes method and classifying the data that will produce what percentage of quantitative data is positive and negative data.

e. Report Preparation

At this point, the report is being prepared as documentation to assist the education and development of others in the process.

III. Research Design

Python is a programming language that has a wide diversity. All you need is the right tools and libraries and you can become a true innovator. Python is extremely easy to read. As an interpreted language (a programming language that doesn't need to be compiled), Python doesn't modify its code to make it computer readable. This language is also a high-level general-purpose programming language. The developers designed it to be the chameleon of the programming world (Muis & Muhammad, 2023). Therefore, this study uses the Python programming language, Google Colab to retrieve data that will be displayed in Microsoft Office Excel and modeled with the help of a library in the Python programming language. The research path carried out in this study can be seen in Figure 1.

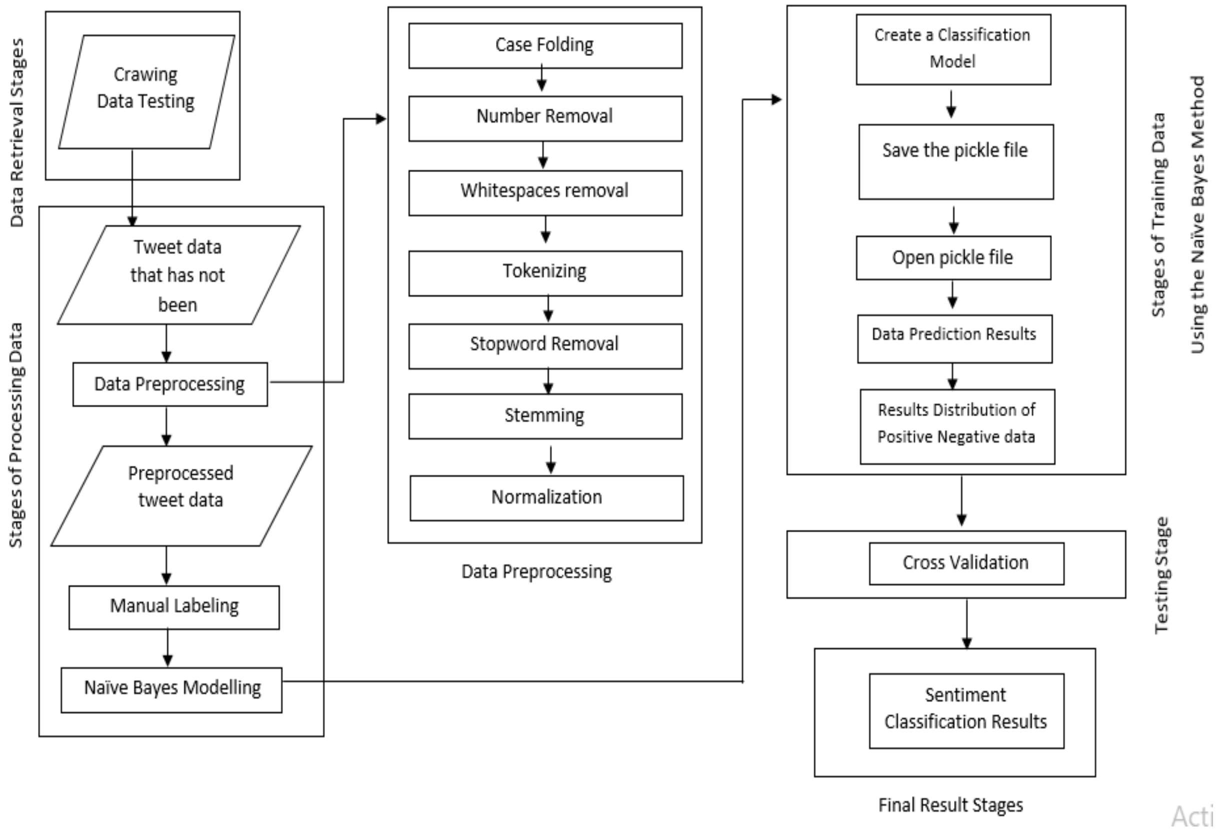

Figure 1 Flow Chart of Research Path

Description of Figure 1 are:

1. **Data Collection Stage**
   Tweet data collection is done by using several libraries available on Google Colab. csv which has a function to read data contained in files with csv format, pandas which has a function to manipulate and clean data and datetime has a function to call operations related to time.

2. **Data Processing Stage**
   Processing data is the process of processing text data that is already available by performing steps to improve text data that has not been repaired.

3. **Data Preprocessing Stage**
   Preprocessing data requires several libraries in Python to help with some tasks. The libraries used for data preprocessing are numpy to support numerical computations or calculations, pandas to prepare data for cleaning, nltk to support natural

4. **Data Training Phase of the Naive Bayes Method**
   Data training is a training process on data using the Naive Bayes Classification method.
   a. Classification Model Making

Followed by making a classification model using variables X and y with available training data.

b. Save File Pickle
The model named pipeline is then saved in the form of a .pickle file so that it can be reopened and used again.

5. **Testing stage**
Testing is a stage to determine the level of validity of the model built at the training stage which is used to predict the label or class from the available test data.

6. **Final Result Stage**
At this stage, sentiment classification results will be carried out in the Classification Report Data Testing.

RESULTS AND DISCUSSION

The tweet data taken is tweets related to the keyword "Kartu Pra Kerja" during the period 01 June 2022 to 05 June 2022 with a total of 6,658 tweets and re-tweet data. To evaluate the categorization model, we used 1000 tweets, 500 of which were categorized as positive or negative. To determine the correctness of the application, it is important to run a confusion matrix test to determine the real and projected values in the data. As can be seen in Table 1, the training data were used to calculate the Confusion Matrix.

Table 1 Confusion Matrix Data Training Results

Actual Class	Prediction Class	
	Positif	Negatif
Positif	90	14
Negatif	13	83

The results of the confusion matrix in Table 1 are TP = 90, TN = 14, FP = 13 and FN = 83. After getting the results from the confusion matrix, it is followed by cross-validation calculations to determine the correctness of the model. Ten repetitions of k-fold cross validation are used to arrive at the correct number. Each fold in k-fold cross validation has its own accuracy value, and the results differ from each other in this regard. The results of the calculation of cross validation 10 fold. The results of the ten rounds of cross validation resulted in an average accuracy score of 0.86 as shown in Table 2.

Table 2. Results of Cross Validation Data Training

Fold	**Accuracy**
Fold 1	81,2 %
Fold 2	85,0 %
Fold 3	81,2 %
Fold 4	91,2 %
Fold 5	87,5 %
Fold 6	82,5 %
Fold 7	90,0 %
Fold 8	86,2 %
Fold 9	88,8 %
Fold 10	87,5 %

Based on table 2. The classification evaluation stage uses testing data as many as 200 new tweets with a positive label of 100 data and a negative label of 100 data. The new data that has been labeled is then calculated for its accuracy value to determine the difference between training data and testing data. From the results of the classification calculation, the average accuracy is quite good, namely 0.874. The results of the classification calculations on the testing data can be seen in Table 3.

Tabel 3. Results of Classification Report Data Testing

Type	Precision	Recall	FI-Score
Negatif	0,94	0,79	0,86
Positif	0,83	0,95	0,89
Accuracy	0,89	0,89	0,88
Macro Average	0,89	0,87	0,87
Weighted Average	0,88	0,88	0,87

And the following are the percentage results of the Accuracy Precision Recall and AUC values as a whole which can be seen in Table 4

Table 4. Percentage Result

Algoritma	Sentimen	*Accuracy*	*Precision*	*Recall*	ACU
Metode Naïve Bayes Classification	Positif	96 %	98%	96%	96 %
	Negatif		92%	96%	

CONCLUSION

Several conclusions can be drawn from the results and debates on sentiment analysis on Twitter social media with the Pre-Employment Card case study is the test results show that the Naïve Bayes Classifier method is successful in classifying sentiment with the highest accuracy value of 96%, the highest precision value of 98%, and the highest recall value of 96%, and AUC of 96%.

Pre-Employment Cards have proven to be highly accepted by Twitter users, according to this study, who look at feelings such as "the benefits, effectiveness, and budget increase of Pre-Employment Cards" in tweets and calculations based on tweet models and data.

REFERENCE

Ananda, F. D., & Pristyanto, Y. (2021). Analisis Sentimen Pengguna Twitter Terhadap Layanan Internet Provider Menggunakan Algoritma Support Vector Machine. *MATRIK : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*. https://doi.org/10.30812/matrik.v20i2.1130

Artha, E. U., Dahlan, A., Informatika, J. T., Teknik, F., Magelang, U. M., Informasi, J. S., & Catur, C. (2018). Klasifikasi Model Percakapan Twitter Mengenai Ujian Nasional. *JPIT*, *03*(01), 121–125.

Buntoro, G. A. (2017). *Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter*. *Integer Journal*.

Fikri, M. I., Sabrila, T. S., & Azhar, Y. (2020). Perbandingan Metode Naïve Bayes dan Support Vector Machine pada Analisis Sentimen Twitter. *SMATIKA JURNAL*. https://doi.org/10.32664/smatika.v10i02.455

Khalimi, A. M. (2020). Perhitungan Confusion Matrix Multi-Class Clasification 3x3.

Laksono, R. A., Sungkono, K. R., Sarno, R., & Wahyuni, C. S. (2019). Sentiment Analysis of Restaurant Customer Reviews on TripAdvisor using Naive Bayes. *2019 12th International Conference on Information & Communication Technology and System (ICTS)*, 49–54.

Muis, A., & Muhammad, F. (2023). Pelatihan Text Mining Menggunakan Bahasa Pemrograman Python. *Abdimas Langkanae*, *3*(1), 36–46.

Pandhu, A., & Diki, W. (2020). Analisa sentimen dan Klasifikasi Komentar Positif Pada Twitter dengan Naïve Bayes Classification. *BRITech (Jurnal Imiah Komputer, Sains Dan Teknologi Terapan*.

Perekonomian, K. (2020). Kumpulan Peraturan Kredit Usaha Rakyat (KUR). *Siaran Pers No. HM.4.6/11/SET.M.EKON.2.3/01/2020*.

Priyanto, A., & Ma'arif, M. R. (2018). Implementasi Web Scrapping dan Text Mining untuk Akuisisi dan Kategorisasi Informasi dari Internet (Studi Kasus: Tutorial Hidroponik). *Indonesian Journal of Information Systems*, *1*(1), 25-33.

Rosdiana, R., Eddy, T., Zawiyah, S., & Muhammad, N. Y. U. (2019). Analisis Sentimen pada Twitter terhadap Pelayanan Pemerintah Kota Makassar. *Proceeding SNTEI*.

Ruhyana, N. (2019). Analisis Sentimen terhadap Penerapan Sistem Plat Nomor Ganjil/Genap pada Twitter dengan Metode Klasifikasi Naive Bayes. *Jurnal IKRA-ITH Informatika*.

Salim, S. S., & Mayary, J. (2020). ANALISIS SENTIMEN PENGGUNA TWITTER TERHADAP DOMPET ELEKTRONIK DENGAN METODE LEXICON BASED DAN K – NEAREST NEIGHBOR. *Jurnal Ilmiah Informatika Komputer*. https://doi.org/10.35760/ik.2020.v25i1.2411

Sanusi, R., Astuti, F. D., & Buryadi, I. Y. (2021). Analisis sentimen pada twitter terhadap program kartu pra kerja dengan recurrent neural network. *JIKO*, *5*(2), 89–99.

Watrianthos, R., Suryadi, S., Irmayani, D., Nasution, M., & Simanjorang, E. F. S. (2019). Sentiment Analysis Of Traveloka App Using Naïve Bayes Classifier Method, *8*(07), 786–788.