# Dhiyauddin Al Ghozi - Application of the Naïve Bayes Algorithm to Analyze LinkedIn Application User Sentiment Regarding Job Vacancies on the Play Store

*by* Layanan Perpustakaan UHAMKA

**Submission date:** 28-Jun-2024 03:38PM (UTC+0700)
**Submission ID:** 2409777661
**File name:** rding_Job_Vacancies_on_the_Play_Store_-_Dhiyauddin_Al_Ghozi.docx (416.86K)
**Word count:** 3184
**Character count:** 17163

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor X, Bulan 2023, Page 999-999
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293

# Application of the Naïve Bayes Algorithm to Analyze LinkedIn Application User Sentiment Regarding Job Vacancies on the Play Store

**Dhiyauddin Al Ghozi[1], Firman Noor Hasan[2,*]**

[1] Faculty of Industrial Technology and Informatics, Informatics Engineering, Universitas Muhammadiyah Prof. Dr. Hamka, Jakarta, Indonesia
Email: [1]ghozialko@gmail.com, [2]firman.noorhasan@uhamka.ac.id
Correspondence Author Email: firman.noorhasan@uhamka.ac.id

**Abstract**– Mobile applications have become an important part, one of which is the LinkedIn application which is a mobile application that focuses on the recruitment process, job search and as a professional networking platform which is now increasingly relevant, especially in Indonesia. The methodology involves data collection, data preprocessing, data labeling, and application of the Naïve Bayes algorithm. Sentiment analysis can be used as a reference to improve the quality of an application and the level of user satisfaction as well as knowing the number of positive and negative sentiments in user feedback. The 999 data obtained were then divided into 60% training data and 40% test data. In this analysis, negative sentiment outweighs positive sentiment, with a total of 539 negative reviews and 460 positive reviews. Based on evaluation using the confusion matrix, accuracy results were 95.74%, precision was 100%, and recall was 91.46%. This research aims to provide insight into the communication and interaction patterns of LinkedIn users in relation to job opportunities and overall sentiment towards the platform.

**Keywords**: Sentiment Analysis; Naive Bayes; LinkedIn; Job Vacancy; Play Store

## 1. INTRODUCTION

A In an era where everything is digital with the rapid growth of technology having an influential impact, one of which is that mobile applications have become an integral part of people's lives [1]. With mobile applications you can access everything quickly and comfortably, users can also connect with the world around them [2]. Mobile applications have made significant contributions in various contexts, including in the fields of education, health and work. A mobile application that can help recruitment and job searching is LinkedIn which provides job information [3].

The LinkedIn application is the third largest online professional networking application that is most widely used in the world [4]. Founded in 2002 by Reid Hoffman and launched in 2003, it is a professional social media platform that allows users to build closeness and professional networks [5]. It is a very important aspect in career management and recruitment to create job opportunities, internships and experiences that strengthen professional relationships. As well as pursuing the skills needed for future career success [6]. Play Store is an official Google application that runs on Android-based mobile devices [7]. According to data from Napoleon Cat, there has been an increase in LinkedIn users in Indonesia by 24.9 million users from January to June 2023. The high number of LinkedIn users makes it easier to access the information needed. However, the availability of a lot of information with uncertainty can cause information overload, where users find it difficult to sort out relevant information [8].

Based on Trading Economy data, Indonesia is one of the countries with the highest unemployment rate in Southeast Asia. In February 2023, the unemployment rate in Indonesia was recorded at 5.45%. According to the Central Statistics Agency (BPS), there are 7.86 million unemployed people in Indonesia in August 2023, a decrease of around 6.77% compared to August 2022 [9]. Even though it has decreased, if we look at the previous year, unemployment in Indonesia has still not recovered to pre-Covid 19 pandemic levels in 2019. This means that there are still many Indonesians who need jobs to support their families.

Previous research has been carried out regarding the application of the Naïve Bayes algorithm in analyzing sentiment. Research conducted by Bobby Kurniadi Widodo, Nur Hafifah Matondang, and Desta Sandya Prasvita regarding the application of the Naïve Bayes algorithm for using the jobstreet application using 1000 datasets resulted in an accuracy value of 96%, precision 98% and recall 94% [10]. In research conducted by Artanti Inez Tanggraeni, and Melkior N Sitokdana regarding reviews of E-Government applications with 674 datasets, it resulted in 89% accuracy, 83% precision, and 87% recall [11]. Other research conducted by Rizky Kurnia Pratama, and Putry Wahyu Setyaningsih analyzed comments on employment using the Naïve Bayes method with 3929 tweet data to produce 97% accuracy, 92% precision and 99% recall [6].

Based on previous research, this research will investigate public opinion on LinkedIn regarding job vacancies and user satisfaction levels by applying the Naïve Bayes algorithm. This research aims to provide an insight into the communication and interaction patterns of LinkedIn users in relation to job and career opportunities


Submitted: **dd/mm/yyyy**; Accepted: **dd/mm/yyyy**; Published: **dd/mm/yyyy**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor X, Bulan 2023, Page 999-999
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293

and overall sentiment towards the platform. In addition, to determine whether there is positive sentiment or negative sentiment in user reviews of the LinkedIn application.

## 2. RESEARCH METHODOLOGY

The research methodology section will be explained systematically as the stages carried out to solve the research problem. Using the Naïve Bayes model, data analysis was carried out as a tool for analyzing sentiment [12]. Figure 1 displays the research methodology flow.

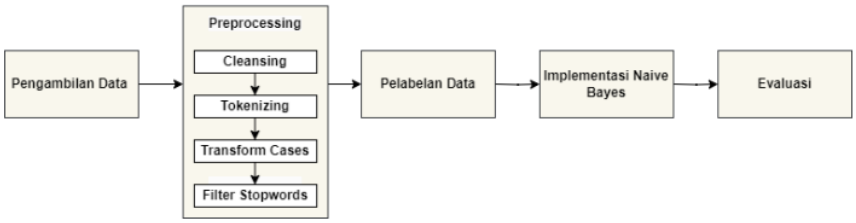



**Figure 1.** Research Flow

The initial stage of the research methodology starts from collecting data using web scrapping techniques, where the data taken is comments from users of the LinkedIn application on the Play Store with the help of Google Colab [13]. The data obtained from the scrapping results in CSV form is continued to the preprocessing stage for processing. The preprocessing stage is the data cleaning stage which includes cleansing, tokenizing, transform cases, and stopwords filter which aims to make it easier for researchers to carry out labeling. After preprocessing, we then carry out labeling. There are two parts of data in this research, namely training data and test data [14]. Manual data labeling carried out on training data is used to train the model before the model carries out classification. Meanwhile, test data is used in the Naïve Bayes implementation process. The stages of the process of implementing Naïve Bayes involve forecasting data and determining whether the data should be classified as positive or negative sentiment based on its similarity to past experiences [15]. The final stage of this research is evaluation, which is to measure the accuracy of the results of the process that has been carried out. By using the confusion matrix equation you can indicate the accuracy, precision and recall values. Below is the calculation formula used in the confusion matrix.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

## 3. RESULT AND DISCUSSION

### 3.1 Data Collection

Data collection was the first stage in this research which was taken using web scrapping techniques using the Python programming language. This data was obtained on November 21, 2023. Web scrapping is a method used to collect information in the form of review data with the aim of extracting information [16]. As shown in Figure 2, the data collection process begins with accessing the Play Store, then going to the LinkedIn application to get the link which will later be entered into Google Colab. After that, the data resulting from the scrapping process is saved in CSV file format and downloaded to proceed to the preprocessing stage.


**Figure 2.** Data Scrapping Process


Submitted: dd/mm/yyyy; Accepted: dd/mm/yyyy; Published: dd/mm/yyyy

**JURNAL MEDIA INFORMATIKA BUDIDARMA**
**Volume 7, Nomor X, Bulan 2023, Page 999-999**
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293

Researchers carried out the process of scraping data from user reviews of the LinkedIn application. Figure 3 shows data scrapping carried out using Google Colab, the first stage is entering 2 libraries, namely numpy and pandas. Then enter the destination link for the LinkedIn application, namely 'com.linkedin.android' and use the latest review category [17]. The data taken was 1000 review data containing user name labels, ratings, dates and reviews. Then the scrapping data is saved in CSV format. After the scrapping process is complete, the CSV file will be downloaded to proceed to the preprocessing stage.

```
[ ]  from google_play_scraper import Sort, reviews

     result, continuation_token = reviews(
         'com.linkedin.android',
         lang='id',
         country='id',
         sort=Sort.NEWEST,
         count=1000,
         filter_score_with=None
     )
```

**Figure 3.** LinkedIn Application Review Data Scrapping

### 3.2 Data Preprocessing

The preprocessing process aims to clean the data so that the data analysis steps can run smoothly. The tool used in the preprocessing process is RapidMiner. At this stage, several stages in preprocessing are carried out, including:

a. The first stage is cleansing. The purpose of this operator is to sterilize documents from components that have nothing to do with document data [18]. Figure 4 shows the cleaning process where the data is tidied up by using the read CSV operator to enter the scrapping data and then connected to the select attributes operator to select the part to be cleaned. Then it is connected with the replace operator to delete parts of the data such as punctuation, characters, hashtags, mathematical symbols, emoticons, and unnecessary symbols ([.?,:!#$+™;*^«âš¾ðŸ‡°@ '¤]) will be omitted.

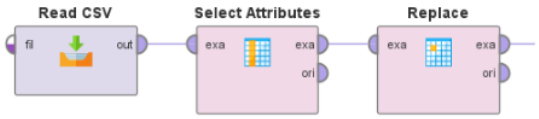



**Figure 4.** Cleansing Stage

Table 1 is the result of the cleansing stage where the data is clean of characters, symbols and punctuation in the review data.

**Table 1.** Cleansing Results

| Before Cleansing | After Cleansing |
|---|---|
| Good for job seekers, keep going | Good for job seekers keep going |
| Very helpful in developing my business. | Very helpful in developing my business. |
| Apk@ Really Helps the Young Generation Who Don't Have a Job... | Apk@ Really Helps the Young Generation Who Don't Have a Job... |

Figure 5 shows the preprocessing operator section which will proceed to the tokenizing, transform cases, and stopwords filter stages after going through the cleansing process as the first stage.


Submitted: **dd/mm/yyyy**; Accepted: **dd/mm/yyyy**; Published: **dd/mm/yyyy**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor X, Bulan 2023, Page 999-999
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293


**Figure 5.** Operators in Preprocessing

b. The second stage is tokenizing. The aim of this operator is to separate or break down words per word in a sentence into individuals [19].

**Table 2.** Tokenizing results

| Before Tokenizing | After Tokenizing |
|---|---|
| Good for job seekers continue | Good, for, job, seekers, continue |
| Very helpful in developing my business | Very, helpful, in, developing, my, business |
| Apk is very helpful for the younger generation who don't have jobs yet | Apk, is very, helpful, for, the younger, generation, who, don't, have, jobs, yet |

c. The third stage is transform cases. The purpose of this operator is to change every word containing uppercase letters in the dataset to all lowercase, using the lower case alternative so that all data is uniformly small in size [14].

**Table 3.** Results of Transform Cases

| Before Transform Cases | After Transform Cases |
|---|---|
| Good, for, job, seekers, continue | good, for, job, seekers, continue |
| Very, helpful, in, developing, my, business | very, helpful, in, developing, my, business |
| Apk, is very, helpful, for, the younger, generation, who, don't, have, jobs, yet | apk, is very, helpful, for, the younger, generation, who, don't, have, jobs, yet |

d. The fourth stage is the stopwords filter. The purpose of this operator is to remove irrelevant or unimportant words that are referenced from the Indonesian stopwords dictionary, such as personal pronouns and conjunctions, which have no sentiment value [20].

**Table 4.** Stopwords Filter Results

| Before Stopwords Filter | After Stopwords Filter |
|---|---|
| good, for, job, seekers, continue | good, job, seekers, continue |
| very, helpful, in, developing, my, business | very, helpful, in, developing, business |
| apk, is very, helpful, for, the younger, generation, who, don't, have, jobs, yet | apk, is very, helpful, the younger, generation, have, jobs, yet |

### 3.3 Data Labeling

After completing the Preprocessing process, the next stage is the labeling process on the dataset. Manual labeling is done by dividing the data into training data and test data. From the scrapping results, 1000 data were obtained, then the number of datasets changed to 999 review data after going through the Preprocessing stage. Researchers divided the data in a ratio of 60:40 for training data and test data [21]. A total of 599 data were used for training data and 400 data were used for test data. Of the 999 data obtained, 599 user comments were manually


Submitted: dd/mm/yyyy; Accepted: dd/mm/yyyy; Published: dd/mm/yyyy

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor X, Bulan 2023, Page 999-999
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293

labeled positive or negative. The remaining 400 user comments were not labeled for testing in Naïve Bayes classification. The purpose of this labeling is so that each dataset has its own positive and negative sentiments to be used as a reference, where positive labels indicate comments containing praise or user satisfaction with the application [22]. Meanwhile, negative labels are comments that contain complaints or user dissatisfaction with the application.

**Table 5.** Sentiment Labeling Results

| Sentiment | Review |
|---|---|
| Positive | It is very helpful for finding work and being able to share it with others in the network |
| Negative | Very difficult to enter login Always returns to the main menu |
| Positive | Very helpful in professional networking |
| Negative | It's still difficult to log in and ask me to choose an image that isn't clear |

### 3.4 Algorithm Implementation

After going through the preprocessing and labeling stages, the next stage is algorithm implementation, using the Naïve Bayes method to develop an algorithm model based on a previously designed model. By using the Naïve Bayes method, you can estimate the probability of similarity between previous experiences and future probabilities [23]. The use of the Naïve Bayes method has shown significant accuracy in handling large amounts of data. Figure 6 shows the process of creating training data. Starting by using the read CSV operator to read the CSV file that has been prepared, then connected with the example filter operator. This operator is set to the option (is not missing) to filter data that is not missing so that later sentiments that are still empty or have not been labeled in the training data will be retrieved and processed [24]. Next, to process the data you need a nominal to text operator so that the document process operator can read and process the data. The process document operator contains the tokenize operator, transform cases, and stopwords filter. Data analysis is then carried out using the Naïve Bayes operator, and the data is stored in stores called the model store and training data store.

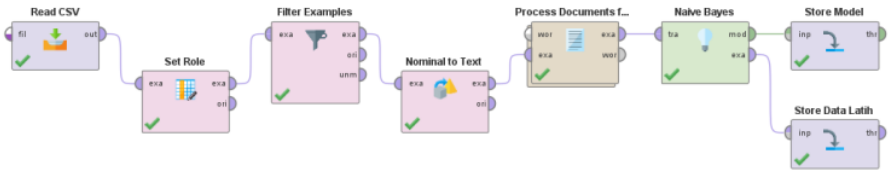
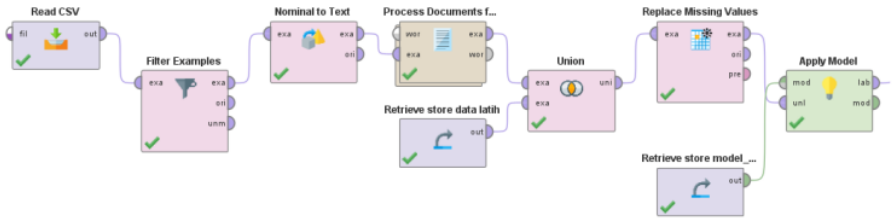



**Figure 6.** Training Data Creation Mechanism

After creating the training data, the next steps for implementing Naïve Bayes are shown in Figure 7. The stages of implementing Naïve Bayes begin with the CSV read operator and then connected with the examples filter operator to eliminate unnamed or missing data. The nominal to text operator and the process document operator have the same explanation as in Figure 6. Then the results of the data analysis in Figure 6, which is called the training data store, are combined together with the process document operator. Next, it is connected with the replace mising values operator, while the data obtained from the analysis which is called the store model is connected with the apply model operator together with the data from the previous analysis or what is called the training data store.





Submitted: dd/mm/yyyy; Accepted: dd/mm/yyyy; Published: dd/mm/yyyy

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor X, Bulan 2023, Page 999-999
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293

**Figure 7.** Naïve Bayes Implementation Mechanism

After going through the stages of preprocessing, labeling, and implementing Naïve Bayes, prediction results are obtained. Figure 8 displays a visualization of the results of implementing Naïve Bayes in the form of a pie chart. From a total of 999 data processed, 460 comments were categorized as positive sentiment and 539 comments were categorized as negative sentiment. The prediction results are based on the research conducted, it can be concluded that user opinions regarding the LinkedIn application are generally negative. This is caused by many users who regret the repeated login verification on the LinkedIn application.

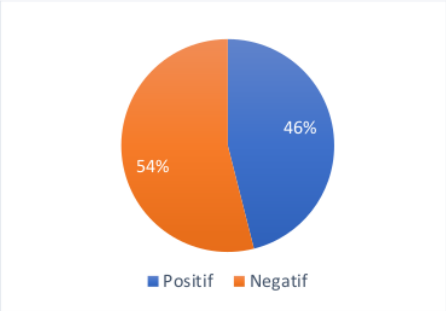



**Figure 8.** Pie chart

Not only that, the results of implementing Naïve Bayes can also be visualized in wordcloud form to find out the words that appear most often and those that are most often used in user reviews. There were 999 review data used in this research, but the researchers only showed 20 words with the highest frequency. The most frequent occurrence is the word 'application' with a total occurrence of 407 words. Due to the many user reviews that regret the repeated login verification on the LinkedIn application. Figure 9 displays a visualization in wordcloud form where the greater the frequency of words appearing, the larger the size of the words in the display format.


**Figure 9.** Wordcloud visualization results

### 3.5 Evaluation

The final stage in this research process is evaluation. The purpose of this evaluation is to determine the results obtained by a model [25]. So that in classifying data we can know the level of accuracy of the Naïve Bayes algorithm. This research applies the confusion matrix equation to determine the results [26]. Figure 10 displays the Confusion matrix which is used to calculate the level of accuracy, precision and recall.

| | | Predicted Class | |
|---|---|---|---|
| | | **Positive** | **Negative** |
| **Actual Class** | **Positive** | **TP** | **FP** |
| | **Negative** | **FN** | **TN** |


Submitted: dd/mm/yyyy; Accepted: dd/mm/yyyy; Published: dd/mm/yyyy

**JURNAL MEDIA INFORMATIKA BUDIDARMA**
**Volume 7, Nomor X, Bulan 2023, Page 999-999**
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293

**Figure 10.** *confusion matrix*

The meaning of the terms used in the confusion matrix is True Positive (TP) if the actual and predicted class are positive. False Positive (FP) if the actual class is negative and the predicted class is positive. False Negative (FN) if the actual class is positive and the predicted class is negative. And True Negative (TN) if the actual and predicted class are negative. Figure 11 shows the various operators used in the confusion matrix calculation process in Rapidminer.

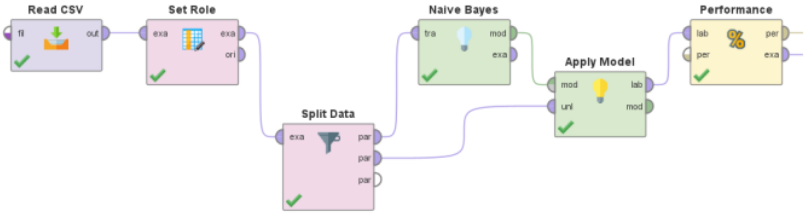



**Figure 11.** Confusion Matrix Mechanism

The confusion matrix mechanism begins with the CSV read operator to enter a file containing prediction data, then the set role operator is connected to read the prediction (sentiment) column as a label. The split data operator is used to split the data with a ratio of 60% to 40% and then connected to the Naïve Bayes operator and the apply model operator. Next, the apply model operator is connected to the performance operator to find out the accuracy results. The results of the confusion matrix process stage produce an accuracy value of 95.74% which is shown in Figure 12.

**accuracy: 95.74%**

| | true Positif | true Negatif | class precision |
|---|---|---|---|
| pred. Positif | 182 | 0 | 100.00% |
| pred. Negatif | 17 | 200 | 92.17% |
| class recall | 91.46% | 100.00% | |

**Figure 12.** Confusion Matrix results

By using the confusion matrix equation, the results of calculating accuracy, precision and recall from sentiment analysis obtained the following results.

$$Accuracy = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} = \frac{182 + 200}{182 + 200 + 0 + 17} = \frac{382}{399} = 0.9573 = 95.74\%$$

Accuracy calculations produce an accuracy of 95.74%, meaning that the model is able to classify the data as a whole at a very good level.

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{182}{182 + 0} = \frac{182}{182} = 1 = 100\%$$

The precision calculation produces a precision of 100%, which means it shows the model is able to identify prediction results correctly.

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{182}{182 + 17} = \frac{182}{199} = 0.9145 = 91.46\%$$

The recall calculation produces a recall of 91.46%, which indicates that the model was successful in finding most of the true positive instances. With high recall, the model tends to be successful in identifying the majority of positive cases.

## 4. CONCLUSION

After conducting research using data from the Play Store using the Naïve Bayes method, we obtained 999 data to be researched by dividing the data, including 599 data used for training data and 400 data used for test data. Of the 999 data processed, 460 reviews were categorized as positive sentiment and 539 reviews were categorized as negative sentiment. Conclusions can be drawn in this research by using the Naïve Bayes method and confusion matrix at the evaluation stage to obtain quite high accuracy values, namely reaching 95.74%, precision of 100%,


Submitted: **dd/mm/yyyy**; Accepted: **dd/mm/yyyy**; Published: **dd/mm/yyyy**

**JURNAL MEDIA INFORMATIKA BUDIDARMA**
**Volume 7, Nomor X, Bulan 2023, Page 999-999**
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293

and recall of 91.46%. The prediction results are based on research conducted that the Naïve Bayes and confusion matrix methods in classifying data have high accuracy. It can be concluded that user sentiment about the LinkedIn application is generally negative. This research shows that the sentiment analysis of LinkedIn application users regarding job vacancies tends to be poor, where negative labels are reviews that contain user dissatisfaction. This is because many users regret the existence of repeated login verification on the LinkedIn application. Suggestions for future researchers are the need to compare several classification methods to find out which method is the best.

## REFERENCES


Submitted: **dd/mm/yyyy**; Accepted: **dd/mm/yyyy**; Published: **dd/mm/yyyy**

# Dhiyauddin Al Ghozi - Application of the Naïve Bayes Algorithm to Analyze LinkedIn Application User Sentiment Regarding Job Vacancies on the Play Store

ORIGINALITY REPORT

| 17% | 16% | 2% | 15% |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | Submitted to Universitas Amikom<br>Student Paper | 13% |
| 2 | Submitted to The University of Manchester<br>Student Paper | 1% |
| 3 | Submitted to University of Hong Kong<br>Student Paper | 1% |
| 4 | jutif.if.unsoed.ac.id<br>Internet Source | 1% |
| 5 | Asrul Sani, Samuel, Djaka Suryadi, Firman Noor Hasan, Ade Davy Wiranata, Siti Aisyah. "Predicting the Success of Garment Sales on Transaction Data using the Classification Method with the Naïve Bayes Algorithm", 2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE), 2023<br>Publication | <1% |
| 6 | Submitted to Deakin University<br>Student Paper | |

| | | |
|---|---|---|
| | | <1% |
| 7 | Firman Noor Hasan, Achmad Sufyan Aziz, Yos Nofendri. "Utilization of Data Mining on MSMEs using FP-Growth Algorithm for Menu Recommendations", MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, 2023<br>Publication | <1% |
| 8 | eprints.uai.ac.id<br>Internet Source | <1% |
| 9 | Sharma, Neelam, and Saurabh Mukherjee. "A Novel Multi-Classifier Layered Approach to Improve Minority Attack Detection in IDS", Procedia Technology, 2012.<br>Publication | <1% |
| 10 | ibtpi.pelitaindonesia.ac.id<br>Internet Source | <1% |
| 11 | publikasi.dinus.ac.id<br>Internet Source | <1% |
| 12 | ejurnal.stmik-budidarma.ac.id<br>Internet Source | <1% |

| | | | |
|---|---|---|---|
| Exclude quotes | Off | Exclude matches | Off |
| Exclude bibliography | Off | | |