OPINION MINING ABOUT PARFUM ON E-COMMERCE BUKALAPAK.COM USING THE NAÏVE BAYES ALGORITHM

Rizal[1], Muhammad Fikry[2], Annisa Helmina[3*]

[1,2,3] Informatics Engineering Study Program
Universitas Malikussaleh
http://www.unimal.ac.id
[1]rizal@unimal.ac.id, [2]muh.fikry@unimal.ac.id, [3*]annisahelmina17@gmail.com

(*) Corresponding Author

Abstrak—*Informasi berperan sangat penting dalam pesatnya perkembangan di dunia. Banyak orang-orang menggunakan media online untuk mencari informasi salah satunya yaitu untuk mengetahui informasi tentang negatif atau positifnya suatu produk di e-commerce berdasarkan komentar-komentar yang ada. Untuk mengetahui klasifikasi dari semua komentar-kometar tersebut membutuhkan waktu yang cukup lama dalam membacanya. Maka untuk mempermudah dari itu semua di buat suatu sistem klasifikasi untuk menentukan klasifikasi komentar. Dalam proses pengklasifikasian ini menggunakan algoritma naive bayes untuk di jadikan solusi dari permasalahan tersebut. Proses dengan algoritma naïve bayes membutuhkan data training yang digunakan sebagai bahan pembelajaran dari sistem. Data training yang digunakan di ambil dari salah satu e-commerce yaitu Bukalapak.com mengenai produk parfum. Mengambil komentar dari Buakalapak.com digunakan tehnik crawling untuk mengambil komentar dari keseluruhan produk. Data training yang di butuhkan dalam sistem ini adalah 1000 komentar yang terdiri dari 500 komentar training positif dan 500 komentar training negatif. Untuk mendapatkan nilai akurasi di butuhkan 300 komentar uji yang terdiri dari 150 komentar uji positif dan 150 komentar uji negatif. Dari hasil pengujian dengan naive bayes tingkat akurasi yang di dapat cukup baik yaitu dengan nilai presisi 96.44%, recall 96.34% dan akurasi 96.33%.*

Kata Kunci: Opinion Mining, Naive Bayes, Text Mining

Abstract—Information plays a very important role in the rapid development of the world. Many people use online media to search for information, one of which is to find out information about the negative or positive of a product in e-commerce based on the comments that exist. To find out the classification of all comments-comet takes quite a long time in reading it. So, to make it easier than that all made a classification system to determine the classification of comments. In this classification process, the Naive Bayes algorithm is used as a solution to the problem. The process with the Naïve Bayes algorithm requires training data which is used as learning material from the system. The training data used is taken from one e-commerce site, Bukalapak.com regarding perfume products. Taking comments from Buakalapak.com used crawling techniques to retrieve comments from the whole product. The training data needed in this system is 1000 comments consisting of 500 positive training comments and 500 negative training comments. To get the accuracy value, it requires 300 test comments consisting of 150 positive test comments and 150 negative test comments. From the results of testing with Naive Bayes, the accuracy rate can be quite good, namely with a precision value of 96.44%, 96.34% recall, and an accuracy of 96.33%.

Keywords: Opinion Mining, Naive Bayes, Text Mining

INTRODUCTION

The use of the internet is quite diverse and can be said to be multifunctional, such as communication, buying and selling, business, exchanging data and information, and others. The internet has become an actual and universal source of information that anyone can use [1]. In the development of the internet business ideas emerge in their implementation using the internet as a means of buying and selling goods and / or commerce through the internet which has other terms namely electronic commerce or e-commerce. E-commerce in practice is to facilitate the sale and purchase of goods or services between business organizations and business organizations with consumers[2]. The usual steps for potential buyers to decide whether or not to buy goods in an e-commerce site is to look at product reviews[3].

Product reviews affect whether or not a product is good [4].

Other people's opinions can be very important information when making choices or decisions[5]. Seeing the number of consumers who write opinions continues to increase, then reading the entire review will take a long time so that if only a few reviews are read, then the evaluation will be biased[6]. Sentiment analysis or opinion mining is a process in presenting information from the processed process of a system used to classify comments into positive or negative comments.[7].This opinion mining research discusses how to analyze sentiments towards e-commerce customer opinions on Bukalapak which are used to determine whether the opinion is positive or negative [8].

Several studies have been conducted in conducting opinion mining on product comment reviews including, tweet clustering in Indonesian language, social media using the Naive Bayes Classifier method[9], a web application for sentiment analysis on product opinion with the Naive Bayes Classifier method[10], sentiment analysis system on product reviews using the Naive Bayes method[11], Sentiment Analysis for Online Shopping Site Assessment Services Using the Naïve Bayes Algorithm[8], the application of sentiment analysis to assess a product on Indonesian Twitter using the Naïve Bayes Classifier and Information Gain method [7], designing a customer comment sentiment analysis system using the Naive Bayes Classifier method [12]and there are many more previous studies that have been done.

Many e-commerce sites include reviews on each product to support the product in its sale. This review can help prospective buyers to make decisions and conclusions about the product to be purchased [13]. Opinions contain valuable information in them, but they are difficult to exploit because of the large and growing numbers and also the lack of knowledge on how to manage these opinions quickly and practically[10]. In classifying a comment, it can be determined by looking at the value of the comment, but not all comments can be determined by just looking at the value. Because there are comments that are not related at all. This is what makes us want to create a comment classifier based on the contents of the comments[9]. Therefore opinion mining is needed to process words or phrases about the review of perfume products to help in concluding the condition of the goods from e-commerce.

In this research, the method used is the Naïve Bayes method which is often called the Naïve Bayes Classifier (NBC). Naïve Bayes Classifier is an algorithm used to find the highest probability value

to classify test data in the most appropriate category [14]. The purpose of this study is to determine the classification of comments into positive or negative classes. Thus in this study, opinions are classified using the Naive Bayes algorithm where comments will be classified into two classes, namely positive and negative classes.

MATERIALS AND METHODS

The data source of this research is qualitative data. The dataset used is perfume comments from Bukalapak e-commerce obtained from https://www.bukalapak.com/ where the comments consist of 1000 comments with 500 positive comments and 500 negative comments for the training data and 300 comments as test data used as a basis for calculating the accuracy of the system with the algorithm used. The dataset to be processed is in the form of a separate set of texts in the form of comments. Comments are obtained by crawling. Comments that will be used for training data and test data will only be comments that correspond to stop list data or term data from comments that have been crawled. Some examples of stop list comments can be seen in Table 1 and Table 2 below:

Table1.Stop List Positive Comments

Feature	One Word	Two Words	Three Words
Harga	Murah	Sangat Murah	Sesuai Isi Dompet
Aroma	Harum	Wangi Semerbak	Membuat Pikiran Tenang
Racikan	Mantap	Sesuai Permintaan	Patut Diacungi Jempol

Table2.Stop ListNegative Comments

Feature	One Word	Two Words	Three Words
Harga	Mahal	Berat Diongkos	Membuat Kantong Kering
Aroma	Busuk	Sangat Menyengat	Membuat Kepala Pusing
Racikan	Jelek	Sangat Jelek	Racikan Asal Asalan

The method chosen in this study is a method based on the suitability of the data with the best text classification method and has been used by several previous researchers namely Naive Bayes[15]. The Naive Bayes algorithm is a classification algorithm based on probabilities in statistics proposed by Thomas Bayes that predicts future opportunities based on past opportunities (Bayes theorem). This method is then combined

with "naive" where the conditions between attributes are independent of each other[16]. The steps in research with the Naive Bayes algorithm can be seen in Figure 1 below:
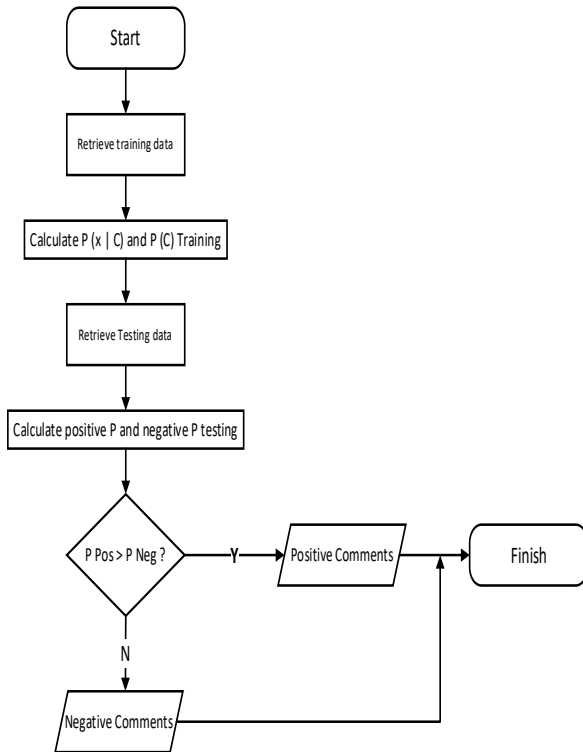

Figure1.Algorithm workflow of N*aïve Bayes*

Naïve Bayes algorithm can be straightened as follows:

$$P(C \mid x) = \frac{P(x \mid C)\, P(C)}{P(x)} \dots\dots\dots\dots\dots\dots(1)$$

Remarks P (x | C) is a conditional probability, P (C) is a prior probability of a class that is the same number of words in all comments, P (x) is the number of all comments in a class. With more general words, the Naive Bayes formula can be given as follows:

$$Posterior = \frac{likelihood\ x\ prior}{evidence} \dots\dots\dots\dots\dots(2)$$

RESULTS AND DISCUSSION

Data that has been crawled will be saved into the database. The next step is to carry out the stages of the process as follows:

1. Data Collecting
Data collection (data collecting) is the taking of a buyer's opinions[3]. Comments for training data

are 1000 and 300 for testing data. Some testing comments can be seen in the following Table 3:

Table 3. Data Testing

No	Comments	Initial Class	Final Class
1	cocok untuk ibadah	Positive	Positive
2	Wanginya enak banget. Hanya kurang tahan lama wanginya	Negative	Positive
3	barang sesuai dengan keinginan	Positive	Positive
4	Wanginya tahan lama tapi tidak suka aromanya tidak lembut. Enakan yg warna kuning.	Negative	Positive
5	saya suka sekali produk ini tidak kecewa saya beli	Positive	Negative
6	Wangi nya enak..	Positive	Positive
7	Aromanya mantaapp...	Positive	Positive
8	Wanginya kurang	Negative	Positive
9	harga murah	Positive	Negative
10	Harga standar,kwalitas lumayan, aroma lumayan wangi..tpi wanginya kok aku kurang suka ya..hehe..brrfungsi sesuai minya wangi????	Negative	Positive
11	wanginya tidak tahan lama	Negative	Negative
12	harum n tahan lama	Positive	Positive
13	Cuma yg warna putih kecewa bau nya bkin pusing dan cepet ilang wanginnya	Negative	Positive
14	jelek ajah barang nya biasa saja	Negative	Negative
15	tidak tahan lama	Negative	Negative
16	Wangi segar	Positive	Positive
17	Wanginya kurang	Negative	Positive
18	wanginya cuman sebentar, satu jam langsung ngilang.	Negative	Positive
19	wanginya hanya hitungan menit	Negative	Positive
20	Lumayan g wangi amat seperti aslinya	Negative	Negative

2. Pre-processing
The pre-processing stage is the initial preparation stage before entering the next process[3].

a. Tokenization
Tokenization is all words and removes punctuation or symbols that are not letters[17], such as " , . /. ;) and others. Examples of tokenization processing results in the comments can be seen in the following Table 4:

Table 4. Tokenization Stage

Raw Comments	*Tokenization*
Kurang harumnya, cepat hilang wanginya.....gak menarik, kecewa	Kurang harumnya cepat hilang wanginya gak menarik kecewa

b.   Stop Word Removal

Stop Word Removal is eliminating common words that do not have the meaning or information needed[17]. An example of the results of processing stopwords removal in the comments can be seen in the following Table 5:

Table 5. Stop Word Removal Stage

Tokenization	Stop Word Removal
Kurang harumnya cepat hilang wanginya gak menarik kecewa	Kurang harumnya cepat hilang wanginya gak menarik kecewa

c.   Stemming

Stemming is one of the processes of changing a token that has a root word to the root, by removing all the prefixes that exist in the token[17]. An example of the results of stemming processing in the comments can be seen in the following Table 6:

Table 6. Stemming Stage

Stop Word Removal	Stemming
Kurang harumnya cepat hilang wanginya gak menarik kecewa	kurang harum cepat hilang wangi tidak tarik kecewa

3.   Classification

This stage is the last in opinion mining. This stage serves to determine the value of sentiment towards a product or service[3]. Examples of the results of classification processing to get the final results in the comments above can be seen in the following Table 7:

Table 7. Stage of Final Results of Test Data

Probability	**Positive**	**Negative**
One Word	0.5249501	1.1317365
Two Words	0	0.1556886
Three Words	0	0.0039920
Average	0.1749834	0.4304724

From the results of the calculations in Table 7, the Negative value is 0.4304724 and the Positive value is 0.1749834. Negative calculation value is higher than the comments that are tested produce a negative classification.

4.   Accuracy Measurement

This last stage aims to test the accuracy of the system with the Naïve Bayes algorithm. Where the accuracy calculation method uses a confusion matrix[3].

Table 8. Class Prediction

Class	Positive	Negative
Positive	148	2
Negative	9	141

Table 9. Confusion matrix

Class	TP	FP	FN	TN
Positive	148	9	2	141
Negative	141	2	9	148

a.   Precision is used to measure the level of accuracy between the information requested by users with the answers given by the system[18].

$$Precision = \frac{TP}{FP+TP} \; x \; 100\% \dots\dots\dots\dots\dots\dots(3)$$

$$Positive = \frac{148}{9+148} X \; 100\% = 94.27\%$$

$$Negative = \frac{141}{2+141} X \; 100\% = 98.6\%$$

b.   Recalled to measure the success of the system in finding back information in the equation[18].

$$Recall = \frac{TP}{FN+TP} \; x \; 100\% \dots\dots\dots\dots\dots\dots(4)$$

$$Positive = \frac{148}{2+148} X \; 100\% = 98.67\%$$
$$Negative = \frac{141}{9+141} X \; 100\% = 94\%$$

c.   Accuracy is useful to know the level of accuracy of the prediction results of class classification against the actual class[10].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \; x \; 100\% \dots\dots\dots(5)$$

$$Positive = \frac{148+141}{148+141+9+2} X \; 100 = \; 96.33\%$$

$$Negative = \frac{141+148}{141+148+2+9} X \; 100 = 96.33\%$$

Table 10 Measurement accusations

Class	*Precision*	*Recall*	*Accuracy*
Positive	94.27%	98.67%	96.33%
Negative	98.60%	94.00%	96.33%
Average	96.44%	96.34%	96.33%

1.   System Design

The Opinion Mining system design with the Naive Bayes algorithm was created using UML (Unified Modelling Language).

a. Usecase Diagram

Figure 2. Usecase diagram of opinion mining System

The use case diagram in Figure 2 above is the work process of an opinion mining system. The context in this diagram has two actors, namely as user and admin. Users can only see the results and details of comment classifications by entering comments on the comment content form and can also see the "Tentang Aplikasi" page and the "Bantuan" page while the admin can manage the system by logging in first.
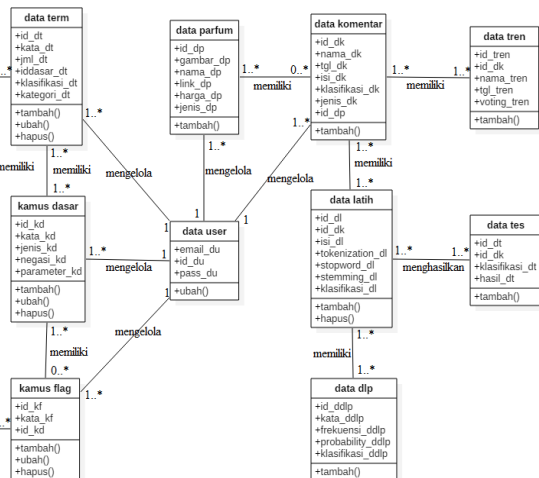
b. *Class Diagram*

Figure 3. Class diagram of the opinion mining system

Figure 3, illustrated in the class diagram there

is a relationship between classes, the explanation is as follows:
1. Relationship between *data_user* and *kamus_dasar*, *kamus_flag*, *data_term*, *data_parfum*, *data_komentar* is that there is only one *user* and that *user* can manage one or many *kamus_dasar*, *kamus_flag*, *data_term*, *data_parfum*, *data_komentar*
2. Relationship between *kamus_dasar* and *data_term* is there are one or many *kamus_dasar* have one or many *data_term*
3. Relationship between *kamus_dasar* and *kamus_flag* is there are one or many *kamus_dasar* have zero or many *kamus_flag*
4. Relationship between *data_komentar* and *data_latih*, *data_tes* there is only one or many *data_komentar* that have one or many *data_latih*, *data_tes*
5. Relationship between *data_parfum* and *data_komentar* there is only one or many *data_parfum* that have zero or many *data_komentar*
6. Relationship between *data_latih* and *data_tren* is that there is only one or many *data_komentar* that produce one or many *data_tren*
7. Relationship between *data_latih* and *data_dlp* there is only one or many *data_latih* that have one or many *data_dlp*
2. Program Implementation
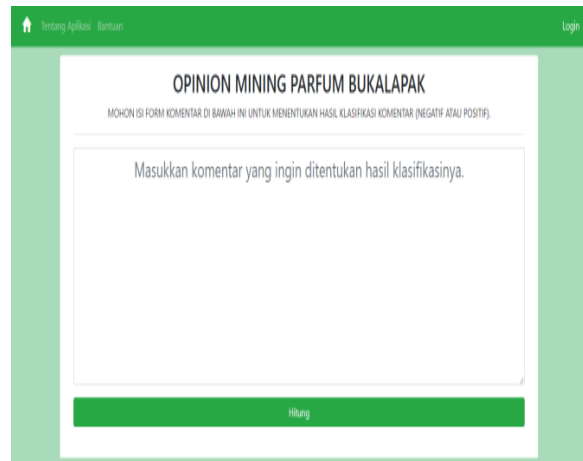a. Display of Front Page

Figure 4 front page

The "Isi Komentar" form in Figure 4 is the form that the user uses to determine the classification of a comment. After the comment is entered, the user can see the classification results and the classification search details of the comments entered.

b. Display of Dashboard page

Figure 5 Display the dashboard page

The "Dashboard" page in Figure 5 is the main page after logging in successfully. On this page, the admin can see the main page of the Opinion Mining application as well as being able to see and update trending perfume scents.
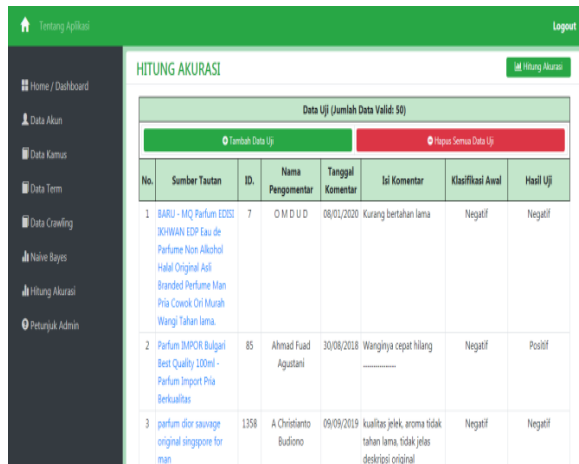
c. Page Count Accuracy

Figure 6 Page count accuracy

The "Hitung Akurasi" page in Figure 6 is a page that contains the calculation of accuracy to see the accuracy of the system in determining the classification of comments. On the page only the admin can access it

CONCLUSION

The conclusions drawn from the results of the analysis and testing conducted on this system include the opinion mining system that is able to classify the comments that exist on the Bukalapak online selling site automatically. Comments for the training data used in this study amounted to 1000 comments, so that if the training data used in learning is large, the classification process is more accurate, but can result in a longer classification process. In the testing phase of comments with Naïve Bayes using 300 test comments resulted in an accuracy of 96.33%.

REFERENCE

[1]   F. N. Zuhri and A. Alamsyah, "Menggunakan Naive Bayes Classifier Di Forum Kaskus Public Sentiment Analysis of Smartfren Brand Using Naive Bayes Classifier on Kaskus Forum," *e-Proceeding Manag.*, vol. 4, no. 1, pp. 242–251, 2017.

[2]   S. Prasetyo and T. Widodo, "Anteseden Kepercayaan Pengguna Pada Penawaran E-Commerce Dan Konsekuensinya Terhadap Niat Beli (Studi Pengguna E- Commerce Provinsi Dki Jakarta) Antecedent of Trust Users on Offering E-Commerce and Its Consequences To the Purchase Intention (Study User," *e-Proceeding Manag.*, vol. 4, no. 2, pp. 1429–1436, 2017.

[3]   Y. Hapsari, M. F. Hidayattullah, and M. Khambali, "Opinion Mining Terhadap Toko Online Di Media Sosial Menggunakan Algoritma Naïve Bayes," *J. Inform. Pengemb. IT*, vol. 03, no. 02, pp. 233–236, 2018.

[4]   S. Ernawati, "Penerapan Particle Swarm Optimization Untuk Seleksi Fitur Pada Analisis Sentimen Review Perusahaan Penjualan Online Menggunakan Naïve Bayes," *J. Evolusi*, vol. 4, no. 1, pp. 45–54, 2016.

[5]   D. Pakpahan and H. Widyastuti, "Aplikasi Opinion Mining dengan Algoritma Naïve Bayes untuk Menilai Berita Online," *J. Integr.*, vol. 6, no. 1, pp. 1–10, 2014.

[6]   D. A. Kristiyanti, "Analisis Sentimen Review Produk Kosmetik Menggunakan Algoritma Support Vector Machine dan Particle Swarm Optimization Sebagai Metode Seleksi Fitur," *Semin. Nas. Inov. dan Tren*, pp. 134–141, 2015.

[7]   A. W. Attabi, L. Muflikhah, and M. A. Fauzi, "Penerapan Analisis Sentimen untuk Menilai Suatu Produk pada Twitter Berbahasa Indonesia dengan Metode Naïve Bayes Classifier dan Information Gain," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 11, pp. 4548–4554, 2018.

[8]   Muljono, D. P. Artanti, A. Syukur, A. Prihandono, and D. R. I. M. Setiadi, "Analisa Sentimen Untuk Penilaian Pelayanan Situs Belanja Online Menggunakan Algoritma Naïve Bayes," *Konf. Nas. Sist. Inf.*, pp. 8–9, 2018.

[9] R. T. Adek and S. Nasution, "Tweet Clustering in Indonesian Language Twitter Social Media using Naive Bayes Classifier Method," *Eurasian J. Anal. Chem.*, vol. 13, no. 6, pp. 277–284, 2018.

[10] S. Hanggara, T. M. Akhriza, and M. Husni, "Aplikasi Web Untuk Analisis Sentimen Pada Opini Produk dengan Metode Naive Bayes," *Semin. Nas. Inov. Dan Apl. Teknol. Di Ind.*, pp. 1–6, 2017.

[11] B. Gunawan, H. S. Pratiwi, and E. E. Pratama, "Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naive Bayes," *J. Edukasi dan Penelit. Inform.*, vol. 4, no. 2, p. 118, Dec. 2018.

[12] E. M. Sipayung, H. Maharani, and I. Zefanya, "Perancangan Sistem Analisis Sentimen Komentar Pelanggan Menggunakan Metode Naive Bayes Classifier," *J. Sist. Inf.*, vol. 8, no. 1, pp. 958–965, 2016.

[13] D. Dwi and J. Santoso, "Multinomial Naïve Bayes Classifier Untuk Menentukan Review Positif Atau Negatif Pelanggan Website Penjualan," *Semin. Nas. "Inovasi dalam Desain dan Teknol.*, pp. 117–122, 2015.

[14] Mihuandayani, E. Feriyanto, Syarham, and Kusrini, "Opinion Mining Pada Komentar Twitter e-KTP Menggunaan Naive Bayes Classifier," *Semin. Nas. Teknol. Inf. dan Multimed*, p. 6, 2018.

[15] Y. T. Arifin, "Komparasi Fitur Seleksi Pada Algoritma Support Vector Machine Untuk Analisis Sentimen Review," *J. Inform.*, vol. 3, no. 2, pp. 191–199, 2016.

[16] A. R. C and Y. Lukito, "Klasifikasi Sentimen Komentar Politik dari Facebook Page Menggunakan Naive Bayes," *JUISI*, vol. 02, no. 02, pp. 26–34, 2016.

[17] R. Wati, "PENERAPAN ALGORITMA NAIVE BAYES DAN PARTICLE SWARM OPTIMIZATION UNTUK KLASIFIKASI BERITA HOAX PADA MEDIA SOSIAL," *JITK (Jurnal Ilmu Pengetah. dan Teknol. Komputer)*, 2020.

[18] R. K. Dinata, Fajriana, Zulfa, and N. Hasdyna, "Klasifikasi Sekolah Menengah Pertama/Sederajat Wilayah Bireuen Menggunakan Algoritma K-Nearest Neighbors Berbasis Web," *CESS (Journal Comput. Eng. Syst. Sci.*, vol. 5, no. February, pp. 33–37, 2020.

Deliberately left blank