Generating SQL Command Syntax Using MySQL Based on Typing Command Sentence

Bonifacius Vicky Indriyono[1], Zudha Pratama[2]

[1,2] *Information System, College of Informatics and Computer Management (STMIK) Kadiri, Indonesia*
[1] bonifaciusvicky@gmail.com (*)
[2] zudhapratama@gmail.com

Abstract— Information retrieval system is a system that is widely used to retrieve information. This research will discuss how the system finds back the information stored in database tables. Tables in the database are arranged to store all forms of data entered by database users so that later the stored data can be used again. Re-accessing the database's information must go through a mechanism known as a database management system (DBMS). One of the most widely used DBMS is MySQL. By using a DBMS, information and data can be manipulated according to user needs. Data manipulation in the database is done in a special language, namely SQL (Structure Query Language). Mastery of SQL commands is an obligation for database users so that the manipulated data can produce the required information. However, many database users still do not understand how the actual SQL command syntax manages and manipulates data into information. This is, of course, very risky if the solution is not immediately sought because it will hinder the process of retrieving information from the data stored in the database. For this problem to be resolved, it is necessary to design a system that can help database users translate their wishes into SQL command syntax. This paper will discuss how a command in Indonesian can be translated into SQL command syntax. The method used to solve this research problem is rule-based. There are two stages in the main process: the pre-processing stage, which consists of a word tokenization process, and a translation stage, including a keyword grouping process. This keyword grouping process consists of the keyword group analysis phase, table and column analysis, identification of SQL commands, and mapping of SQL commands. From all stages that have been passed and testing of 7 scenarios with ten (10) commands for each scenario, the accuracy is 81.42%. The inaccuracy in the testing process is more a problem of displaying data from two or more tables, for example, using the join table command. This problem can be addressed by adding new rules for the use of table joins.

Keywords— Database, DBMS, MySQL, SQL, Command Sentence.

I. INTRODUCTION

The database collects organized, organized and managed, and appropriately stored data in computer storage media. The data stored in the database can be processed or manipulated using application software to become useful information. The database collects information stored systematically to simplify and speed up the information retrieval process [1]. According to Connoly [2], a database is a collection of data designed logically to complement an organization's information needs. For the information stored in the database to be conveyed to database users, the data must be managed or manipulated using a unique system mechanism. The system is a Database Management System. Some database systems have metadata that is part of systematic data management in a DBMS (Database Management System) [3].

DBMS is a software system that allows users to define, create, maintain, and control access to databases [2]. Laudon [4] explains that DBMS is software that enables an organization to centralize data, manage data, and provide access to data stored by application programs. According to Indra [5], DBMS is software used to control the creation of maintenance, processing, and use of large databases. DBMS is also designed in such a way to make it easier to manipulate data. DBMS had become a significant role or key and standard part in supporting a company. By using a DBMS, database users can manipulate data into useful information. There are many DBMSs, one of which is MySQL. MySQL is one of the most well-known DBMS. MySQL functions as a

Relational Database Management System (RDBMS) and is open so that MySQL is considered suitable for demonstrating the database replication process [6]. In a relational database, there is the term constraint. In this case, Constraints are significant to pay attention to because if they are not, then the resulting data will not be valid and cannot fulfill the elements as data for testing [7]. The process of manipulating data in a database requires special tools recognized by the DBMS in use. This tool is known as SQL (Structured Query Language).

SQL is a structured language used to access data stored in databases. SQL is also a standard language that can be applied in various existing databases to be easier to use even though it is easy to move between databases, even though moving from one database to another [8]. According to Junus [9], SQL is a standard language for interacting with databases by operating common commands in the database. Based on its function, SQL is classified into three types, namely DDL (*Data Definition Language*), DML (*Data Manipulation Language*), and DCL (*Data Control Language*). In the context of its use in database management, the CRUD (Create, Read, Update and Delete) concept of the three SQL groups is used to describe the function of a data storage area [10]. SQL commands (or commonly referred to as SQL statements) are issued to the DBMS, which will then run the SQL command and will display the results (if any) [11]. However, many database users cannot use SQL commands to get certain information from the data stored in database tables. This is due to a lack of understanding and the user's ability to use SQL. Sometimes,

user confusion is also caused by determining what SQL commands should be used to get the expected information.

This study aims to help database users get the expected information even though their understanding of SQL is still lacking. Also, another goal of this research is how database users can quickly generate commands typed in Indonesian into SQL command syntax form. According to Khalek [12], to promptly generate or get SQL command syntax, you can use a query creation framework called the Random Query Generator (RQG). The working concept is to use SQL grammar as the basis for generating queries.

However, to achieve the stated objectives in this study, an application was created that could translate user commands into SQL command syntax. In this application, users can type the command in Indonesian, and then the application will display the actual SQL command syntax. Suppose a user wants to display married employees. In this application, the user can type the command in Indonesian, and then the application will display the actual SQL command syntax. For example, a user wants to display married employees. Then in the application, the user will simply type the command: Show employees whose marital status is Married. The application will automatically display the following SQL command syntax from this command: Select * from tbl_karyawan where *stskawin* = 'Married'.

II. RESEARCH METHODOLOGY

The research method is generally defined as a scientific way to obtain data with specific purposes and uses [13]. In this sub-section, the methods used in this research will be explained.

A. Action Research Method

In conducting this research, the method used is a case study. The case study research method is one type of research that can provide answers to several main ideas about a particular event. The case study research method allows the investigation of certain events, circumstances, or social conditions [14].

The steps of the researcher in the case study method are shown in Figure 1. The steps taken by the researcher in this case study can be described as follows:

- Determining Research Topic: Researchers carry out selecting topics by considering specific objectives according to the case to be researched and a solution sought. In this case, the case to be investigated is how to create a system that can help database users use SQL commands based on typing command sentences in Indonesian.
- Determining and Defining Questions: At this stage, the researcher arranges questions on the problem under study and the ultimate goal of this research. This research's main problem is that there are still many database users who do not understand SQL. The question is formulated about a solution to make it easier for database users to write SQL command syntax. From the questions that have been formulated, then the researcher

analyzes these questions and looks for answers about what methods are used to answer questions and implement them into the system.

- Determining Research Design: At this stage, the researcher will evaluate the design and instruments used in the unit's study or subject under study. The researcher will evaluate a situation that must be resolved by determining whether to use a single or multiple-case design in the research and selecting the appropriate instrument with the research question.
- Collecting Research Data: At this stage, the collection of data relevant to the topic under study and the formulation of questions that have been made is carried out. Researchers used data collection methods through observation and literature study in scientific journals, books that discuss similar cases, and other references.
- Data Analysis and Validation: In this step, from the data collected, the researcher will sort out which data is most suitable and can be tested into the system to be built.
- Creating Documentation: The final step in this case study method is to generate documentation. Documentation is made by the researcher in a written report on the research's overall implementation.
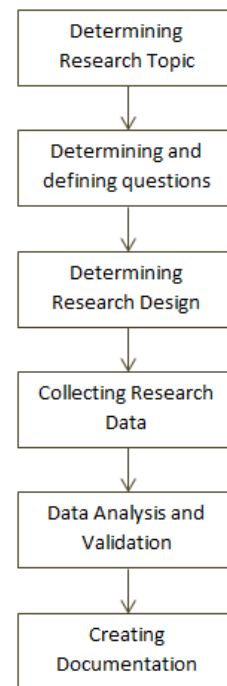

Figure 1. Action research steps

B. Collecting Data Method

Data collection methods are techniques or methods used to collect data. The method refers to how its use can be demonstrated through questionnaires, interviews, observations, tests, documentation, etc. According to Subarsimi [15], data collection methods are defined as researchers' procedures in collecting research data. Data collection methods are the main thing in research activities because the research aims to get data.

In this research, data collection was carried out secondarily, namely through literature study. A literature study is carried out by studying scientific work results from previous researchers, both from journals and proceedings that have themes relevant to current research. In addition to scientific literature, researchers also studied books that discuss the Query language and SQL generator techniques. From the results of this literature review, it is hoped that researchers will gain in-depth knowledge of methods for making SQL generators effectively based on users' commands to reduce the level of difficulty of users in using SQL commands.

C. Analysis Data Method

The data analysis method is part of the implementation of research activities where the data that has been collected by the researcher is processed and used as an answer to the identification or formulation of the problems obtained.

The data analysis method used in the implementation of this research is qualitative analysis. The qualitative analysis method is a procedure intended to understand the symptoms or problems faced by research subjects, including behavior, actions, and research subjects' perceptions.

In this method, the researcher can formulate several issues: the inaccuracy of using SQL commands to execute what the user wants. This inaccuracy is since the user does not even master SQL command forms in displaying the required information.

D. Research Flow

The research flow is designed so that the research implementation stages are more focused so that the results will be as expected. The flow of research activities is shown in Figure 2.
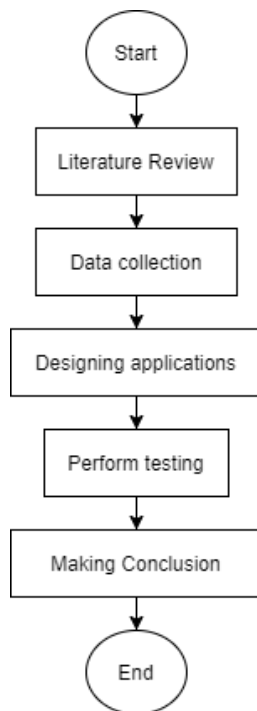
Figure 2. Research flow

Based on Figure 2 above, the system flow can be described as follows :
Start

- Literature review: the researchers looking for research that discusses SQL syntax and looking for a study that examines text processing
- Data collection: the researchers search for the dictionary table, formulate keyword groups and formulate SQL syntax pattern phrases for displaying data. Including look for examples of Indonesian command sentences to display data
- Designing application: the researchers decide the SQL syntax pattern rules for displaying data then implementing a programming language
- Perform testing: The researchers tested the application with examples of Indonesian command sentences from previous data collection results. Then, make adjustments to the application until the results are close to the designed SQL syntax pattern rules, including Record test results.
- Concluding: the researchers determine conclusions from the test results regarding the suitability of the system in the application with the expected results

III. RESULT AND DISCUSSION

This section describes the system's flow, pre-processing text analysis, tokenizing process analysis, and keyword grouping process up to the application implementation stage.

A. System Flow

The resulting system from this research is focused on displaying the SQL command syntax using MySQL based on typing the command sentence by the user. In general, the running of the system is shown in Figure 3.
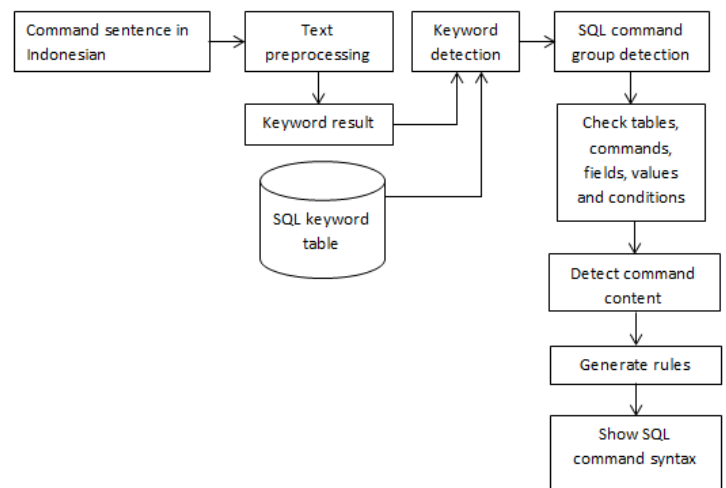
Figure 3. System flow

Based on Figure 3 above, the system flow can be described as follows :
- The user types the command using Indonesian.
- After the command is typed, the next step is the pre-processing text.

3

- From the pre-processing text process, the result is a list of keywords.
- The keyword list is then detected to match with the SQL keyword table.
- The matched keyword is used to define SQL commands, whether it belongs to the DDL or DML group.
- After getting the group from the SQL command, the next process is to check whether it is included in the table, command, field, value, or condition category.
- The next step after the command is detected to check the contents of the user's command and match it with the SQL command group.
- In the last stage, the system will display the SQL command syntax

B. Pre-processing Text

Pre-processing text is the first step in preparing input data, namely command sentences in Indonesian. This input is carried out the first time before the translation activities are carried out. Pre-processing text is a necessary process that needs to be done to obtain excellent accuracy results [9]. The pre-processing text generally consists of the case folding, filtering, word tokenizing, stemming, and stopword removal stages. The pre-processing text stages are shown in Figure 4.
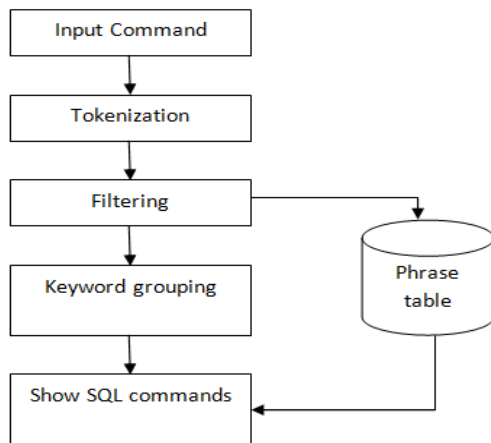

Figure 4. Pre-processing text flow

The pre-processing text stage based on Figure 4 above can be explained as follows :

- Commands in Indonesian are entered into the system.
- From the entered command, the tokenization process is carried out where this process is a process where a long command sentence is cut into words.
- Furthermore, the filter process is carried out from the tokenization results. This filtering process is used to compare words with the existing table of phrases. Its purpose is to find out whether the words have a special symbol. For example, the same as being a sign (=), less than being a sign (<), and so on.
- After the filtering process, then proceed with the keyword grouping process. This process aims to determine whether the command entered belongs to the

DDL or DML group, whether it also includes a condition, clause, or condition value.
- The final step is to display the SQL command syntax based on typing the command sentence in Indonesian.

C. The Tokenizing Process

Tokenizing is defined as a process of dividing a sentence into words (tokens) using a separator parameter, namely a blank space ("). With the parameters, a sentence can be used to collect words (tokens), which can be in numbers or words [16]. The resulting words will be processed at a later stage.

The tokenizing process is also applied in the research results system this time. This process is intended to analyze each word that has been separated from the command sentence entered by the user. An example of the result of the tokenizing process in Table I.

TABLE I
EXAMPLE OF THE TOKENIZING PROCESS

Command sentence	The result of the tokenizing process	
	Word index order	The word
Show all students who have a male gender	0	Show
	1	All
	2	Students
	3	Who
	4	Have
	5	Male
	6	Gender

D. Keyword Grouping Process

In this keyword grouping process, the word tokenizing results are detected into what SQL command group and translated into SQL command syntax form according to the SQL command group. This keyword grouping is divided into four processes: keyword group analysis, table and column analysis, SQL command analysis, and SQL command syntax mapping.

1) Keyword group analysis: Each root word resulting from the pre-processing process, namely tokenizing, will be analyzed to determine the type of root word included in the type of command, value, condition, or other types. An overview of the keyword group analysis is shown in Figure 5.
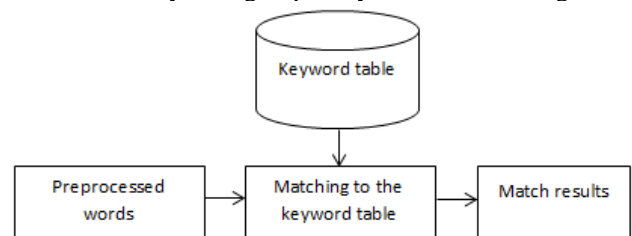

Figure 5. Keyword group analysis process

A group of keywords is generated from the process of keyword analysis, as in Table II.

TABLE II.
KEYWORD ANALYSIS RESULTS

Basic word	Keyword group
Show	
All	Command
Student	
Have	Value

Basic word	Keyword group
Male	
Gender	Condition

2) Analyze tables and columns: Each base word that is not a keyword will be used in the table and column analysis process. Its function is to automatically know each word's type so that it can be used for the next process. The flow of the table and column analysis process is shown in Figure 6 below, and the results of the table and column analysis process are shown in Table III.
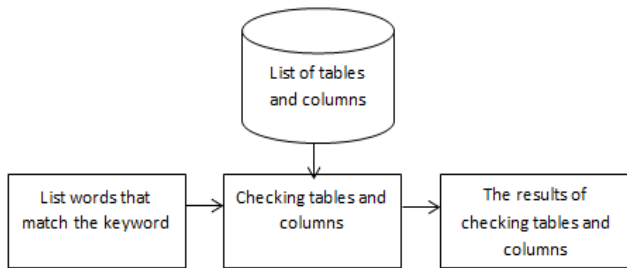

Figure 6. Table and column analysis process.

TABLE III.
RESULTS OF THE TABLE AND COLUMN ANALYSIS

Basic word	Keyword group
Show	
All	Command
Student	Table
Have	Condition
Male	Value
Gender	Column

3) Identify SQL commands: This identification process is carried out to determine whether the user's command is included in the DDL or DML group. After knowing the SQL language group, it is continued to identify each word of the command with the SQL command keyword. Does it include: tables, fields, commands, conditions, or values. If everything is known correctly, the command sentence that is typed is compared with the data in the database so that the SQL command can be known [17]. The identification process flow is shown in Figure 7, and examples of the identification process results in Table IV.

TABLE IV.
AN EXAMPLE OF THE RESULTS OF THE IDENTIFICATION PROCESS

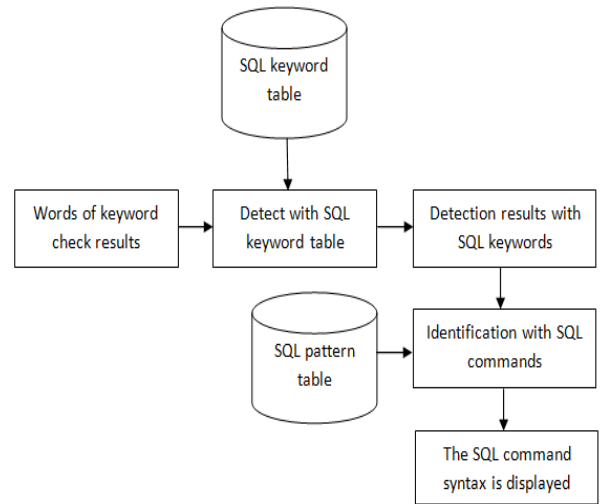| Basic word	Keyword group	SQL command identify
Show		
All	**Command**	**Select**
Student	Table	*Mahasiswa*
Have	Condition	Where
Male	Value	*Kelamin*
Gender	Column	*Pria*

Figure 7. Identification process flow

4) SQL command syntax mapping: At this stage, mapping the SQL command syntax is based on identifying the SQL command. From this process, the SQL command syntax will be obtained according to the user's commands. An example of the results of this process in Table V.

TABLE V.
EXAMPLES OF SQL COMMAND SYNTAX MAPPING RESULTS

Basic word	SQL command identify	Keyword group	SQL command syntax
Show		Command	
All	Select		Select * from
Student	*Mahasiswa*	Table	*mahasiswa*
Have	*Where*	Condition	where
Male	*Kelamin*	Value	*kelamin='Pria'*
Gender	*Pria*	Column	

E. System Testing Analysis

This section will explain the test scenario to determine the level of accuracy of the system being built. There are 7 test scenarios, with each scenario having 10 commands in Indonesian, which will be tested for the SQL command syntax accuracy. Examples of test scenarios, input commands, and expected SQL syntax results are shown in Table VI.

TABLE VI.
EXAMPLE OF THE TESTING PROCESS

Scenario number	Scenario name	Example input command	Expected SQL syntax results
1	Mention the table name directly	show student data	SELECT * FROM t_student
2	Mention the name of the table accompanied by a column, amounting to one column	show a list of student names	SELECT t_student.name FROM t_student

Scenario number	Scenario name	Example input command	Expected SQL syntax results
3	State the name of the table accompanied by a column of two or more columns	show student number and student name	SELECT t_student.studentnumber, t_student.name FROM t_student
4	Mention the name of the table with the filter, the number of one filter	show the data of students whose college status has passed	SELECT * FROM t_student WHERE collegestatus = 'passed'
5	Specifies the name of the table with the filter, two or more filters	show the data of information system study program students who graduated in 2019	SELECT * FROM t_student WHERE studyprogramme = 'information system' AND graduatedin = '2019'
6	Mention table name along with columns and filters	show a list of names of students from the information systems study program	SELECT t_student.name FROM t_student WHERE studyprogramme = 'information system'
7	State the names of two different tables	show a list of students and lecturers of the information system study program	SELECT * FROM t_student, t_lecturers WHERE studyprogramme = 'information system'

Table VI shows examples of tests in 7 scenarios for getting SQL command syntax. From the overall trials for all scenarios, the test results are obtained as in table VII. The total accuracy value in table VII is obtained from the number of correct divided by the total data multiplied by 100%. Judging from the test results as in table VII, in general, the system can perform its functions quite well in terms of the accuracy value obtained.

TABLE VII.
TEST RESULTS TABLE

Scenario Number	The number of commands tested	Total Correct	Total False	Accuracy (%)
1	10	10	0	100
2	10	10	0	100
3	10	10	0	100
4	10	8	2	80
5	10	8	2	80
6	10	7	3	70
7	10	4	6	40
Total	70	57	13	81,42

IV. CONCLUSION

Based on the results of the trials and accuracy tests that have been carried out. The system built in this study has a level of functionality that is already running properly. After being tested with several scenarios and commands in Indonesian, the system could translate user commands into SQL command syntax with a reasonably good accuracy value of 81.42%. The accuracy value is quite good, but there are still things that are not suitable, namely when the system is asked to display data from a different table, for example, a command that uses a join table.

The future work is the SQL command dictionary table in this application is still incomplete. If there is an input command that does not match the table, the results are not optimal. This requires the process of identifying commands and automatically recording them in the basic SQL command dictionary table. This application's system can still detect and repair if there are commands that emphasize a condition initially. For example: Show the highest score of student test results. We need to add more capabilities to detect and display SQL command syntax with any conditions. We need to add a special rule to handle commands that require the system to display data from different tables. as an example, command with table join

The application produced from this study functions to display the SQL command syntax based on the command sentence typed by the user. For further development, the application's deficiencies will be fixed to have a function that runs well in any condition. Also, the application will be developed using voice input commands.

REFERENCE

[1] Suhartati and Y.D. Atma, "Simple Query Optimization for Query Speeds on a Database Server", *METIK Journal,* Vol. 1 No. 1, p. 14-17, Jun. 2017.
[2] T. Connolly, *Database Systems A Practical Approach to Design, Implementation, and Management Fifth Edition.* Boston: Pearson Education, 2010.
[3] E. Tungadi, "Form Generator Utilizing MySQL Metadata", *National Seminar on Communication and Informatics (SNKI),* p. 77-80, Jun. 2015.
[4] K.C. Laudon and J.P. Laudon, *Management Information Systems: Managing the Digital Firm, 11th Edition.* New Jersey: Prentice-Hall. 2010.
[5] I. Warman and R. Ramdaniansyah, "Comparative Analysis of Query Database Management System (DBMS) Performance Between Mysql 5.7.16 And Mariadb 10.1", *TEKNOIF Journal,* Vol. 6 No. 1, p. 32-41, Apr. 2018.
[6] H. Yuliansyah, "Designing Mysql Database Replication With Security Mechanisms Using SSL Encryption", *Journal of Informatics,* Vol. 8 No. 1, p. 826-836, Jan. 2014.
[7] V. Natali and P. Alfadian, "Specific Domain Analysis and Design Language for Data Generator on Relational Database", *JUMANJI Journal,* Vol. 3 No. 1, p. 64-73, April. 2019.
[8] R. Pamungkas, "Query Optimization in My SQL Database Using an Index", Research: Journal of Computer, information systems, & technology management, Vol. 1 No. 2, p. 27-31, Apr. 2018.
[9] J. Sinuraya, "Data Search Methods Using Hash Join Query and Nested Join Query", *Journal of Teknovasi,* Vol. 4 No. 1, p. 42-50**,** Apr. 2017.
[10] M. Fikry, et al. "Generating Web Forms Based on SQL Metadata and W3C specifications", *CoreIT Journal*, Vol. 3 No. 2, p. 63-68, Dec. 2017.

[11]   A.K. Jaiswal, "SQL Query Generator for Natural Language", *International Journal of Research In Science & Engineering,* Vol. 3 No. 2, p. 190-196, March. 2017.

[12]   S.A. Khalek and S. Khurshid, "Automated SQL Query Generation for Systematic Testing of Database Engines", *ASE 2010: The 25th IEEE/ACM International Conference on Automated Software Engineering,* p. 329-332, Sept. 2010.

[13]   Sugiyono, *Educational Research Methods with Quantitative Approaches, Qualitative, and R & D*. Bandung: Alfabeta. 2013.

[14]   D.J. Hodgetts and E.E. Stolte," Case‑based Research in Community and Social Psychology: Introduction to the Special Issue", *Journal of Community & Applied Social Psychology,* Vol. 22 No. 5, p. 379-389, Aug. 2012

[15]   S. Arikunto, *Research Procedure A practical approach.* Jakarta: Rineka Cipta. 2010.

[16]   K.K. Purnamasari and I.S. Suwardi, "Rule-based Part of Speech Tagger for Indonesian Language", in *Proc.* IOP Conf. Series: Materials Science and Engineering, 2018, Vol. 407, p. 1-4.

[17]   D. Ratnasari, D. B. Qur'ani, and A. Apriani, "Sistem Informasi Pencarian Tempat Kos Berbasis Android," *Inf. J. Ilm. Bid. Teknol. Inf. dan Komun.*, vol. 3, no. 1, pp. 32–45, 2018.