ICIC Express Letters
Volume **15**, Number **10**, October **2021**

ICIC International ©2021 ISSN 1881-803X
pp. **1083–1091**

UTILIZING NATURAL LANGUAGE PROCESSING IN CASE-BASED REASONING FOR DIAGNOSING AND MANAGING SCHIZOPHRENIA DISORDER

Sri Mulyana[1], Sri Hartati[1], Retantyo Wardoyo[1] and Subandi[2]

[1]Department of Computer Science and Electronics
Faculty of Math and Natural Sciences
[2]Faculty of Psychology
Universitas Gadjah Mada
North Sekip, Bulaksumur, Yogyakarta 55281, Indonesia
{ smulyana; shartati; rw; subandi }@ugm.ac.id

Received December 2020; accepted March 2021

Abstract. *Some methods have been developed to help diagnose mental disorders and provide treatment. These methods are developed as an alternative to solve the problems of imbalance between mental health services and the number of psychiatrists/psychologists, for example, CBR systems, and expert systems. In most health services, the initial examination of patients with mental disorders is performed by non-specialist medical personnel. At a certain stage when the patient is unable to express the symptoms, the examining officer is required to state the conditions in daily language. Based on this condition, the paramedics will carry out the next diagnosis. In handling of case-based reasoning assisted diagnoses, the use of natural language as an expression of the patient's condition is not an input commonly used in the system. Hence that the natural language processing model becomes symptoms is needed. In this research, natural language processing was developed to produce symptoms according to the patient's condition. The results of this natural language processing will be inputted into a case-based reasoning system. The result of the natural language experiment using 124 data, shows a precision level of 88% and a recall of 67% for the original natural language text. For normalized natural language text, it yields a precision of 92%, and a recall of 78%. This provides input to a case-based reasoning system for diagnosing the type of schizophrenia disorder and its treatment.*
Keywords: Case-based reasoning, Natural language processing, Symptoms, Schizophrenia

1. **Introduction.** Indonesia has several challenges in the health sector, including mental health. Furthermore, the basic health research conducted by the Indonesian Ministry of Health recorded that the prevalence of schizophrenic mental disorders was about 6.7 per mile. The ratio of available psychiatrists to the population in Indonesia is still far from the conditions required by the World Health Organization. Furthermore, there are limited mental health facilities in various regions in the country, and many patients suffering from mental illness are not being properly cared for. However, the rapid development of technology and information could be an alternative in overcoming this imbalance, especially in the field of artificial intelligence. Case-Based Reasoning (CBR) has become a successful technique for knowledge-based systems in many domains.

The ability of health workers to perform anamnesis on patients suffering from mental disorders has not been effective, compared to mental illness specialists. Therefore, inexperienced health workers could describe the conditions experienced by mental patients in simple natural language. In addition, a Natural Language Processing (NLP) based system could be developed to process this natural language to produce symptoms of mental disorders.

DOI: 10.24507/icicel.15.10.1083

2. **Related Work.** Natural Language Processing (NLP) has been widely implemented and has provided support in the field of artificial intelligence, including in the Clinical Decision Support Systems (CDSS) field [1]. Many researchers have applied the NLP technique for input texts processing in the health sector. Stocker et al. developed a text processor in German language which was used to identify risk factors for events that do not comply with operational standards. The NLP technique used was syntax-based keyword matching and the results were used to support patients safety and reduce the malpractice incidence [2]. Anderson et al. developed a text processor based on the clinical reports of high-risk patients using a statistical analysis-based syntactic technique. This method was used to detect clinical events in monitoring the successful treatment of patients at risk [3].

Furthermore, many other NLP techniques had also been implemented in the medical profession. Matheny et al. implemented NLP techniques and rules-based algorithms in English narrative text processors. The method was developed to detect symptoms in various medical conditions [4]. Fung et al. have also developed a narrative text processor in English on drug labels and indications using the medical identification concept with Meta-Map to determine the safety and quality of prescribed drugs [5].

Automatic understanding of natural language problems is a long-standing challenge research problem in automatic solving. Gan and Yu proposed models understanding of geometry questions as a problem of relation extraction, instead of as the problem of semantic understanding of natural language [6].

Research related to CBR has also experienced very rapid developments, especially in the medical field. Chakraborty et al. developed CBR systems for the detection of swine flu called SFDA (Swine Flu Diagnostic Assistant) [7]. Mulyana et al. have developed a CBR system to detect mood disorders [8]. Mulyana and Ilham have also developed a CBR system for post-accident patient treatment [9].

In this research, a natural language processing method was developed from the medical records of mental patients to produce symptoms that will be processed into a case-based reasoning system to help diagnose the type of schizophrenia disorder and its treatment. Mulyana et al. have developed a model based on NLP for processing narrative text of medical records [10].

Schizophrenia is a major psychiatric disorder that affects the perception, thoughts, and behavior of an individual. Though clear awareness and intellectual abilities are usually preserved, certain cognitive deficits may develop later due to the condition [11]. The symptoms of schizophrenia are broadly divided into two groups, namely positive and negative. Positive symptoms include delusions, hallucinations, mental confusion, restlessness, and strange or hostile behavior, and negative symptoms include dull or flat moods, withdrawal or isolation from associations, lack of emotional contact (quiet, difficult to talk), passive, apathy or indifferent, difficulty thinking abstractly, and loss of impulse or initiative.

Schizophrenia diagnosis begins with the Diagnostic and Statistical Manual of Mental Disorders (DSM), namely DSM-III, DSM-IV and DSM-IV-TR. The following are types of schizophrenia according to the DSM-IV-TR classification, and the diagnosis was based on the dominant symptoms observed: paranoid type, disorganized type, catatonic type, residual type, undifferentiated type, and schizophrenia simplex [12].

3. **Methods.** A Natural Language Processing (NLP) based system was developed to process natural language and produce symptoms as inputs into a case-based reasoning system which would be used to help diagnose the types of schizophrenic disorders and its treatment. This research focuses on the use of natural language processing, and the CBR process which works as a complementary system that utilizes the NLP process output.

3.1. **Building a corpus of symptoms and traits.** The corpus built was based on the patient symptoms and description from medical record and was specific for only schizophrenic disorders. Furthermore, each term in the description was manually categorized into one of the following three classes:

1) class symptoms (symbolized by <g></g>);
2) class traits (symbolized by <e></e>);
3) other class (not symbolized, meaning that it will automatically be assigned to all words other than the two categories above).

The corpus was built using the HTML format to facilitate the parsing process. An example of a sentence in the corpus includes:

```
the patient <g>pacing</g>, <g>talking does not connect</g>, <e>difficult
</e> <g>sleeping</g>, <e>sometimes</e> <g>cry alone</g>
```

In the process of forming the corpus, words that were placed in one type of category were written in an element that begins and ends with a corresponding tag. Some examples of the corpus arrangement results on schizophrenia disorders are shown in Table 1.

Table 1. Example of composing a corpus

No.	Narrative text	Labeling result text
1	These last 4 days were difficult taking medication, relapses in the last 2 weeks, sleeping difficulty, pacing, talking loud and rambling, overeating, lack of self-care	These last 4 days <e>was difficult</e> <g>taking medication</g>, relapses in the last 2 weeks <g>sleeping</g> <e>difficulty </e>, <g>pacing</g>, <g>talking loud </g>, and <g>rambling</g>, <g>eating </g> <e>over</e>, <g>taking care of self </g> <e>less</e>
2	throwing things at home, getting angry, screaming, the patient wandering, difficult eating since 2 days ago showering constantly	<g>throwing things at home</g>, <g>angry</g>, <g>screaming</g>, the patients <g>wandering</g>, <e>difficult </e> <g>eating</g> since 2 days ago, <g>showering</g> <e>constantly</e>

3.2. **The NLP (Natural Language Processing).** The components of natural language processing for producing the schizophrenic type symptoms are shown in Figure 1.

Every natural language text input goes through the following processes, namely preprocessing, labeling with NER (Named Entity Recognition), and pattern matching. Preprocessing begins with sentence segmentation which involves dividing the input text into sentences.

For each sentence, case-folding would be carried out by replacing all capital letters with non-capital letters. Furthermore, this is followed by a cleansing process which would be carried out by marking the beginning and end of the sentence with "ESC", changing the abbreviated words, removing spaces between hyphens, adding spaces between dots or commas and doing the tokenization process, and marking each word in the sentence as tokens.

The next process is to label each token with NER. The three entities defined are symptom (g), trait (e), and others (o). A token recognized in the symptom corpus was labeled with "g" and a token recognized on the trait corpus was labeled "e"; otherwise, it was labeled with "o". Furthermore, the last process was pattern matching, which involves finding suitable symptoms with the following steps:

1) Grouping words based on labeling results by combining words that have the same label sequentially;

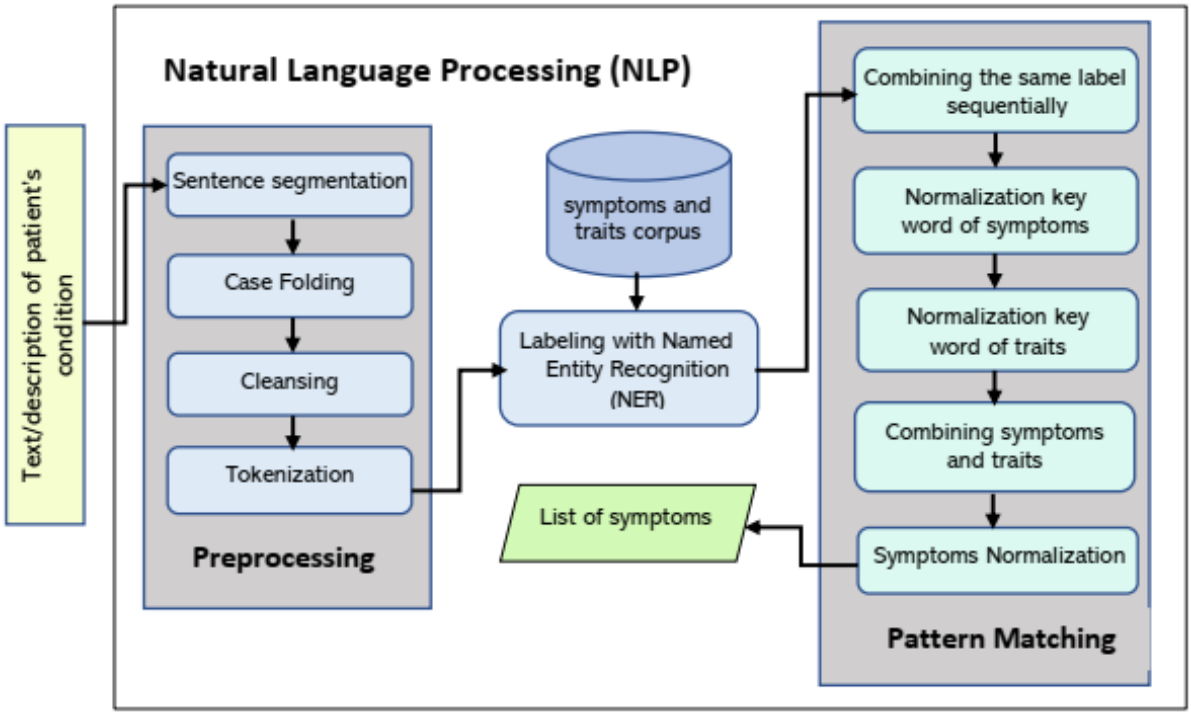


FIGURE 1. NLP process components

2) Determining the type of symptom based on keywords;
3) Combining the corresponding types of "trait" and "symptom" to determine the symptoms.

3.3. **Case-Based Reasoning (CBR) process.** Case-based reasoning is the process of solving new problems based on the solutions of similar past problems. This process will select a case with the greatest similarity level above the specified threshold, that is the value determined by the expert as the boundary of a case having an acceptable level of similarity. The case-based reasoning process sequence for determining the type of schizophrenia disorder and its treatment is shown in Figure 2.

Based on Figure 2, it was revealed that the CBR process gets input in the form of a list of symptoms obtained from the results of the NLP process, and additional symptoms from physical and psychiatric examinations. The next process was to search for the type of schizophrenia disorder and its treatment by calculating the level of similarity between new problem and stored cases. Furthermore, case with the greatest degree of similarity with values greater than or equal to the threshold value will be used as a solution to new problems and this process is commonly referred to as reuse.

4. **Result and Discussion.** The initial stage of processing the input text is by building a symptom sentence pattern and applying the regular expression. It includes adding the first and last signs of the string according to the symptom sentence pattern found in the medical record text.

Based on the natural language processing scheme in Figure 1, the process begin with sentence segmentation by dividing paragraph into sentences, if the input text consists of more than one sentence. Furthermore, words that use period as an abbreviation, would be maintained based on the sentence in the corpus. After the above procedures have been carried out, the next step would be to perform the cleansing process, through the following stages:

1) Marking the beginning and end of the sentence with "ESC";
2) Changing the abbreviated words, for example, "ps" is changed to "patient";

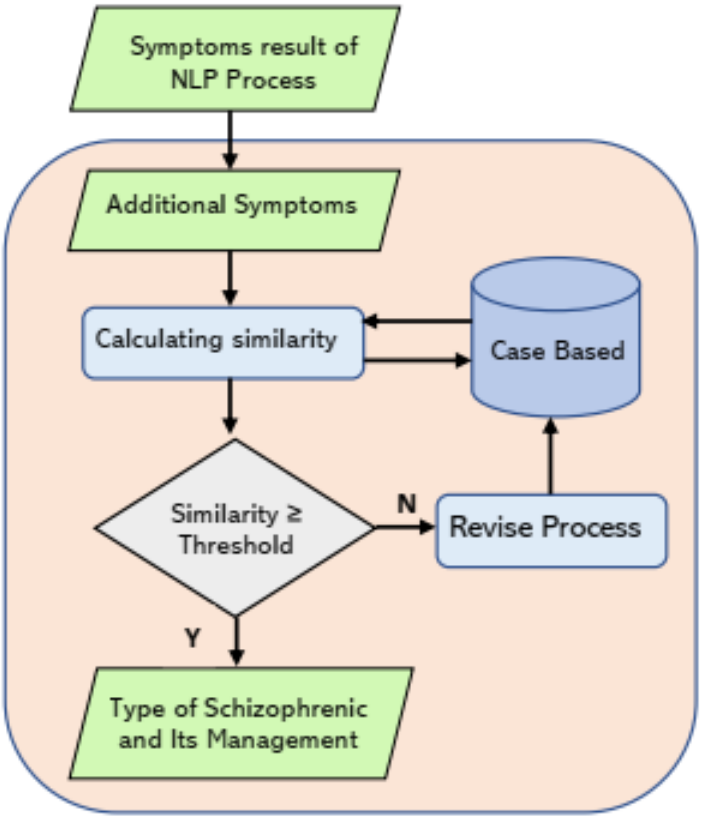


FIGURE 2. Case-Based Reasoning (CBR) process schematic

3) Removing spaces between hyphens;
4) Deleting numbers.

For example, the following input sentence was being read:

`Psn is often angry, has no appetite to drink`

After the cleansing process, the following sentence was obtained:

`ESC patient is often angry, has no appetite to drink ESC`

The next step is labeling using the Named Entity Recognition (NER), by marking each word with a symptom (g), trait (e), or other (o). From the results of cleaning the sentence above, the following NER labeling results were obtained:

ESC	patients	often	angry	,	has	no	appetite	to	drink	ESC
o	o	e	g	o	e	e	e	o	g	o

The trait and symptoms of sequential positions and having the same type will be combined into one keyword, for example, "has" (e), "no" (e) and "appetite" (e) combined into "has no appetite" (e), therefore,

ESC	(patients)	(often)	(angry)	,	(has no appetite)	to	(drink)	ESC
o	o	e	g	o	e	o	g	o

The next step was defining the keyword "symptom" based on the rules of regular expression. Some examples of the keyword expression rule pattern for "symptoms" are shown in Table 2.

The same was applied with the keyword "properties" based on the rules of regular expression as shown in Table 3.

The results of the normalization of "symptoms" and "traits" include

ESC	(patients)	(+)	(angry)	,	(-)	(appetite to drink)	ESC
o	o		g	o		g	o

The next step was to combine the traits and symptoms in a sequence, for example, "+" (e) "angry" (g) are combined into "+ angry". Furthermore, the final step was to

TABLE 2. Example of "symptom" keyword expression rule pattern

No.	Keywords	Regular expression pattern used
1	suicide	"(kill | hang |) yourself", "plunge"
2	talking	"talking $", "speaking $", "^ communication"
3	suspicious	"suspicious", "felt husband", "felt wife", "accused"
4	connect	"^ connect $", "^ obviously $", "focus"
5	confused	"^ confused $",
6	silent	"^ silent $", "^ blank $", "^ shy $", "daydreaming $"
7	emotion	"^ emotion (al)? $"
8	shaking	"^ (body)? shaking $"

TABLE 3. Expression rule patterns for normalizing the keyword "trait"

No.	Keywords	Regex pattern used
1	Positive (symbolized +)	"^ often $", "^ easy $", "^ excess (\\-more)?", "^ routine $", "^ good $", "^ many $", "^ (continu | continuous) $", "^ like $"
2	Negative (symbolized −)	"hard $", "no $", "reduced $", "sometimes $", "rarely $", "refused $", "could not $", "did not want $", "could not $", "little $", "sometimes not $", "less $", "never $" "decreased $", "old $", "\\ (\\ − \\)", "lazy", "bizaree", "cannot", "ugly", "limited", "irregular", "displeased", "disturbed"
3	Neutral (symbolized 0)	"^ can be $", "^ want $", "^ usual $", "^ regular $", "^ normal $", "\\ (\\ + \\)"

TABLE 4. Rule pattern for determining symptom names and types

Symptom code	Symptom name	Regex pattern used
G-25	fright	"afraid"
G-26	difficult in eating	"\− (eat | drink)"
G-27	normal eating	"0 (eat | drink)"
G-28	lazy to move	"\− activity"
G-29	normal shower	"0 shower"
G-30	excessive shower	"\+ shower"
G-31	less shower	"\− shower"
G-32	often angry	"(\+)? angry"

determine the symptom code and name. Several patterns of rules for determining the names and types of symptoms are shown in Table 4.

Based on the rule pattern for determining the names and types of symptoms in the above table, the pattern of "+ angry" was expressed as G-32: often angry, and "− drink" was expressed as G-26: difficult in eating. The process was repeated until all input sentences were processed. In order to determine the performance of the text processing process using NLP, 124 data sets were tested, namely text that contains the description of the patient's condition as stated on the medical record. There were 2 types of text being tested, namely the original text (without making any changes) and the normalized text (make changes to the original text, adjusted to the standard term symptoms of schizophrenia, without changing the meaning). Furthermore, each text that has been normalized was labeled manually in the form of a symptom code according to the list in Table 4 which was used as a reference for calculating the precision. Some examples of the original and normalized texts, and symptom codes are shown in Table 5.

Table 5. Some examples of original and normalized text

No.	Original text	Normalized text	Symptom code
1	Talks to himself sometimes does not clear has difficulty sleeping does not want to eat gets angry laughs throws things into the well	talking to himself, unclear, difficulty sleeping, refusing to eat, getting angry, laughing, throwing things into the well	G-6 G-8 G-51 G-31 G-32 G-49 G-42
2	Talks alone, sometimes clear and sometimes not does not want to drink medicine for 3 days has difficulty sleeping has difficulty bathing eating and drinking gets angry breaks things and burns	talking to himself, sometimes unclear, refusing to take medicine, difficulty sleeping, difficulty taking a shower, difficulty eating or drinking, getting angry, breaking things	G-6 G-8 G-54 G-51 G-31 G-26 G-34 G-42

Table 6. Example of the calculation results of precision and recall level

No.	Symptoms	Original text			Normalized text		
		Symptom results	Precision	Recall	Symptom results	Precision	Recall
1	G-6 G-8 G-51 G-31 G-32 G-49 G-42	G-6 G-8 G-51 G-26	0.75	0.43	G-6 G-8 G-51 G-26	0.75	0.43
2	G-6 G-8 G-54 G-51 G-31 G-26 G-34 G-42	G-6 G-49 G-52 G-41 G-4	0.2	0.125	G-6 G-49 G-51 G-41	0.5	0.25

Based on the example in Table 5, the results of the precision and recall levels obtained after testing are shown in Table 6.

The symptoms produced through text processing with NLP were compared with others that were produced through manual labeling, which was followed by calculating the level of precision and recall [13]. For example, A represents the symptom set on the target, while that of B is as a result of text processing. The precision and recall levels are calculated as follows:

$$Precision = \frac{|A \cap B|}{|B|} \qquad Recall = \frac{|A \cap B|}{|A|}$$

The following are explanations of the calculated results in table above.

Case No-2

A = {G-6, G-8, G-54, G-51, G-31, G-26, G-34, G-42}

$|A| = 8$

Original text:

B = {G-6, G-49, G-52, G-41, G-4}

$|B| = 5$

$A \cap B$ = {G-6}

$|A \cap B| = 1$

$$Precision = \frac{|A \cap B|}{|B|} = \frac{1}{5} = 0.2$$

$$Recall = \frac{|A \cap B|}{|A|} = \frac{1}{8} = 0.125$$

Normalized text:

B = {G-6, G-49, G-51, G-41}

$|B| = 4$

$A \cap B$ = {G-6, G-51}

$|A \cap B| = 2$

$$Precision = \frac{|A \cap B|}{|B|} = \frac{2}{4} = 0.5$$

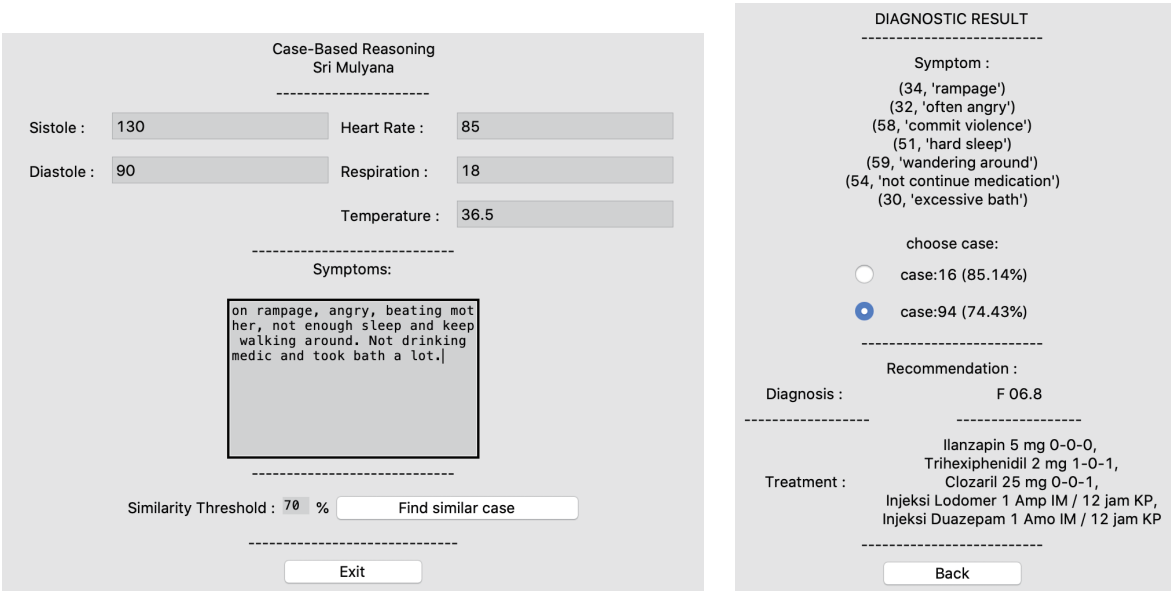
$$Recall = \frac{|A \cap B|}{|A|} = \frac{2}{8} = 0.25$$

TABLE 7. Average of precision and recall levels

Original text		Normalized text	
Precision	**Recall**	**Precision**	**Recall**
0.88	0.67	0.92	0.78

The test results of the 124 data set, based on the average value of precision and recall levels are shown in Table 7.

The development of a case-based reasoning system begins with building a case base from the results obtained from the data collection that have been carried out and stored with JSON extension. The following is an example of implementing a case-based reasoning system. For example, there are new problems, namely: a patient with blood pressure 130/90, pulse 85, respiration 18, and temperature 36.5. After being processed, the symptoms experienced by the patient were obtained, namely: G-34: rampage, G-32: often angry, G-58: commit violence, G-51: hard sleepy, G-59: wandering around, G-54: not continue medication and G-30: excessive bath. Based on these symptoms, and with a threshold of 70%, the CBR system calculates the level of similarity to cases stored on a case basis. The case with the highest level of similarity at 85.14% was case number 16. Therefore, it can be recommended as a solution to new problems with diagnosis and treatment as shown in Figure 3.



(a) (b)

FIGURE 3. (a) Display of input window, (b) diagnostic result

5. **Conclusions and Future Work.** Based on the results of testing for NLP to produce symptoms as inputs into a case-based reasoning system, to help diagnose the types of schizophrenia disorders and their treatment, it is concluded that

1) The NLP system succeeds in processing the sentences of patient complaints by mental disorders patients and extracting them into mental disorders symptoms. These results become input to the CBR system. The NLP system test result provides a precision level of 88% and a recall of 67% for the native language. In addition, for normalized natural language, it has a precision of 92% and a recall of 78%.

2) The CBR system can determine the diagnosis as a mental disorder. A case which has the greatest similarity value of the specified threshold value will be selected. This is a diagnostic result of CBR. The solution to this case is a psychiatric disorder treatment.

In the future, the research can be developed which focuses on collecting the real cases of schizophrenic disorders and its management. This can be done in collaboration with psychiatrist/psychologist, so that the results can be used to assist in diagnosing the real cases.

REFERENCES

[1] J. A. Reyes-Ortiz, B. A. González-Beltrán and L. Gallardo-López, Clinical decision support systems: A survey of NLP-based approaches from unstructured data, *The 26th International Workshop on Database and Expert Systems Applications (DEXA)*, pp.163-167, 2015.

[2] C. Stocker, L. M. Marzi, C. Matula, J. Schantl, G. Prohaska, A. Brabenetz and A. Holzinger, Enhancing patient safety through human-computer information retrieval on the example of German-speaking surgical reports, *The 25th International Workshop on Database and Expert Systems Applications (DEXA)*, Munich, Germany, pp.216-220, 2014.

[3] H. D. Anderson, W. D. Pace, E. Brandt, R. D. Nielsen, R. R. Allen, A. M. Libby and R. J. Valuck, Monitoring suicidal patients in primary care using electronic health records, *The Journal of the American Board of Family Medicine*, vol.28, no.1, pp.65-71, 2015.

[4] M. E. Matheny, F. FitzHenry, T. J. K. Speroff, M. L. G. Griffith, E. E. Vasilevskis and S. H. Brown, Detection of infectious symptoms from VA emergency department and primary care clinical documentation, *International Journal of Medical Informatics*, vol.81, no.3, pp.143-156, 2012.

[5] K. W. Fung, C. S. Jao and D. Demner-Fushman, Extracting drug indication information from structured product labels using natural language processing, *Journal of the American Medical Informatics Association*, vol.20, no.3, pp.482-488, 2013.

[6] W. Gan and X. Yu, Automatic understanding and formalization of natural language geometry problems using syntax-semantics models, *International Journal of Innovative Computing, Information and Control*, vol.14, no.1, pp.83-98, 2018.

[7] B. Chakraborty, S. I. Srinivas, P. Sood, V. Nabhi and D. Ghosh, Case-based reasoning methodology for diagnosis of Swine Flu, *IEEE GCC Confererence and Exhibition (GCC)*, Dubai, United Arab Emirates, pp.132-135, 2011.

[8] S. Mulyana, S. Hartati, R. Wardoyo and E. Winarko, Case-based reasoning with input text processing to diagnose mood [affective] disorders, *International Journal of Advanced Research in Artificial Intelligence (IJARAI)*, vol.4, no.9, 2015.

[9] S. Mulyana and S. Ilham, The determination of the action toward the patient's psychological theraphy in the post-accident using case-based reasoning, *Indonesia Journal of Computing and Cybernetics Systems (IJCCS)*, vol.11, no.1, pp.11-20, 2018.

[10] S. Mulyana, S. Hartati, R. Wardoyo and Subandi, A processing model using natural language processing (NLP) for narrative text of medical record for producing symptoms of mental disorders, *2019 4th International Conference on Informatics and Computing (ICIC)*, vol.1, no.1, pp.1-6, 2020.

[11] B. J. Sadock and V. A. Sadock, Mood disorder, in *Kaplan & Sadock's Synopsis of Psychiatry: Behavioral Sciences/Clinical Psychiatry*, 10th Edition, Wolters Kluwer Phiadelphia, Lippincott Williams & Wilkins, 2007.

[12] R. Maslim, *Pocket Book: Mental Disorders Diagnosis Brief Reference from PPDGJ – III and DSM-5*, Master Thesis, Department of Mental Health, Faculty of Medicine, Universitas Atmajaya, Jakarta, 2013.

[13] G. Kowalski, *Information Retrieval Architechture and Algorithm*, Springer, New York, USA, 2011.