2016 IEEE 6th International Conference on System Engineering and Technology (ICSET)
Oktober 3-4, 2016 Bandung – Indonesia

Scenes Categorization based on Appears Objects Probability

Marzuki[1], Egi Muhamad Hidayat[2], Rinaldi Munir[3], Ary Setijadi P[4] ,Carmadi Machbub[5]

School of Electrical Engineering and Informatics, Institut Teknologi Bandung
Bandung, Indonesia

[1]marzuki@lskk.ee.itb.ac.id
[2]egi.hidayat@ lskk.ee.itb.ac.id
[3]rinaldi-m@stei.itb.ac.id
[4]asetijadi@lskk.ee.itb.ac.id
[5]carmadi@lskk.ee.itb.ac.id

***Abstract*—Automation of visual perception on the machine is the automation of the interpretation of the data generated by the camera sensor like humans. Some automated systems such as the robot and intelligent vehicle relies heavily on an understanding of the environment. The most fundamental understanding of the visual environment is the Machine's ability to identify categories of scenery is still a lot of uncertainty. The uncertainty of the categorization of the scene appears because of the difficulty categories indicated in complex environments, the layout of objects and different scene in the same category. We present a model approach Distribution of the chance appearance of objects and classify them to obtain the semantic meaning of a scene. From the experiments conducted, the model shows high accuracy and using a dataset SUN908 this approach is an effective approach to explore the knowledge-based on the scenes label using large dataset.**

***Keywords*— scene category, machine perception, visual perception, scene understanding, object probability**

I. Introduction

Humans can identify many entities in-surrounding areas easily, despite the fact that the entity that visits can vary from different angles, in different sizes and scale or when the entity is at the layout of the unusual (tilted, rotated and so on), even entities it can be recognized even when partially obstructed from scene. Understanding of scenery is a product of the meaning of relationships between entities around, be it objects, layout and actions that may or are being carried out by the entity. The process for human understanding, studied in psychology and cognition, the results of the meaning process in the field of psychology known as perception. In the field of computer vision called machine perception, which results from the process of finding relationships and identify the meaning of the scene on the image or video that also specifically called visual perception.

Visual pattern recognition through the device's camera technology allows to perform various tasks, such as object recognition, target recognition, navigation, understand and manipulate things. Advances in technology dramatically camera causing it to become one of the sensors of choice for automation.

This automation is tough on fundamental issues such as (a) the arrangement of the sensor used, (b) the interpretation of the data generated by sensors and (c) the level of efficiency of implementation and system execution [1], [2].

Scene Understanding is a fundamental problem in Computer vision. The Scene is referred to in this paper is a place where humans can move in it, or a place that people can navigate, which is generally represented by the pictures. The dynamics incredibly diverse in the real-world causes great scenes need to be classified so that restrictions on activities and navigation is possible can be categorized by the machine properly.

Research developments in the field of scene of the rapid categorization, the general approach taken in the development of classification based on the scenes of datasets and attributes that exist as well as the recognition and classification algorithms through feature extraction, segmentation, objects labelling and others. Development dataset which is phenomenal as well [3]–[6] named SUN908. Xiao and his colleagues believe the research on the scene understanding have been restricted by the limited scope of the database that is used at the time. Existing databases that do not represent the scenes of the various categories as a whole. While the standard database for object category contains hundreds of objects of different classes.

Xiao and his colleagues published a paper back in 2012, behind them is a 3D phenomenon. They assume a system that can automatically understand 3D scenes just by looking not only requires the ability to extract 3D information from the image, but also to handle a wide variety of different environments. To overcome the various scenepoints in the scenes categorization, they have introduced the image database 360^{o} panoramic scenepoint. Furthermore, with the same database, [3] developed a taxonomy of attribute scene through human description. By using a sample of 397 categories of 908 categories to evaluate a variety of state-of-the-art Algorithm, then for the sake of research in the field of the scene understanding, they publish a hierarchical dataset SUN908 on the link *http://vision.princeton.edu/projects/2010/SUN/* and http://*groups.csail.mit.edu/vision/SUN/*.

978-1-5090-5088-8/16/$31.00 ©2016 IEEE

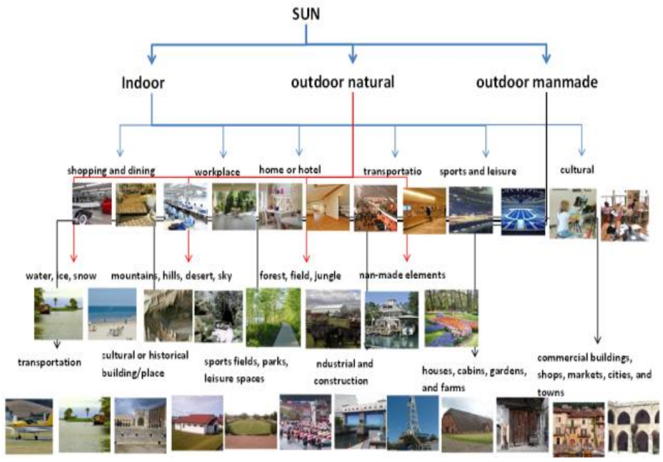
SUN
Indoor
outdoor natural
outdoor manmade
shopping and dining
workplace
home or hotel
transportatio
sports and leisure
cultural
water, ice, snow
mountains, hills, desert, sky
forest, field, jungle
man-made elements
transportation
cultural or historical building/place
sports fields, parks, leisure spaces
ndustrial and construction
houses, cabins, gardens, and farms
commercial buildings, shops, markets, cities, and towns

Fig 1. Hierarchy of Scene Category

II. RELATED WORK

The approach used to resolve the uncertainty in determining the scene category, this paper was inspired by some of the approaches.

A. *Scenes Categorization SUN908 Dataset*

SUN908 is a database of the scene category was first developed by researchers [4] to map objects and scenery category in 2010. SUN908 had 130.519 images were grouped into 899 categories. The development of this database is intended to overcome the limitations of the database which can be used for testing the algorithm in computer vision. For the initial test used 397 categories using algorithms developed by many researchers to recognition scene. In the Year 2012 dataset is expanding into 16.873 pictures and labeled SUN2012 or known by SUN908 which is the identification number of categories that they develop are 908 categories.

First destination database development are trying to map out a variety of different scenes at once a function of the spectacle. The second objective is to identify all environmental and places considerable importance that has a unique identity and build the most complete dataset with the name of SUN (Scene Understanding). Further measures how accurately people can do categories of images are scene and measure the performance of recognition algorithms developed many researchers.

Subsequent developments in 2014, was composed of taxonomic [6] from this database to be easily implemented on an integrated data using electronic dictionaries WordNet selected 70,000 words together to describe the objects associated with a dataset. Currently the database SUN has 908 categories, 131.067 pictures, 313 884 objects that have been segmented manually and has 4479 categories of objects that have been defined globally.

B. Perception in Psychology

The main problem of perception in psychology is how people interpretation the meaning of the information received through the sensor. Associated with visual perception is a process called pattern recognition. Such as object recognition, recognition of action or movement of objects and visual understanding the psychological perception has been studied by Gestalt perception.

Gestalt theory is closely associated with Max Wertheimer (1880-1943), Wolfgang Köhler (1887-1967), and Kurt Koffka (1886-1941). The third founder of the school of psychology Berlin in Germany have formulated "**The laws of grouping**" and formulate a process approach Gestalt perception becomes Bottom-Up and Top-Down Proces (Fig. 2) [7]. The term bottom-up also called data-driven that is using the data ever known (memory). In this approach begins with the perception of pieces of information from the scene and combine in various ways to form a new meaning.

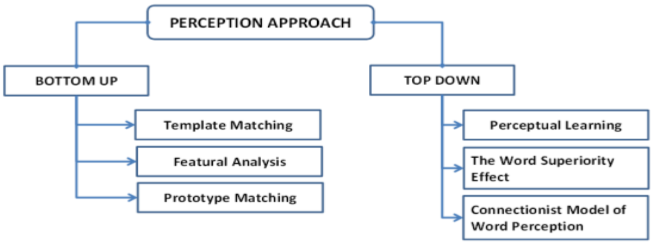

Fig 2. Perception Gestalt Approach, [7]

In this paper used Gestalt perception approach is data-driven approach or a bottom-up Perception. Approach to find the meaning of a set of information, this approach is divided into three approaches, namely template matching, featural analysis and prototype matching.

Computer vision is the science template matching definition is the common name used on a technique to measure the part of the image or to determine whether the two images are the same or not [8]. While the analysis is the process of finding features characteristic of an image or object in the image. To implement both approaches, in this paper using the analysis feature is used inference largest value (maximum) that identifies a set of specific space (equations 8 and 9).

Prototype matching model is the process of perception at the time sensor received information will be followed up by comparing that information with information previously stored in the form of a prototype. This approach does not require a definite similarity between the stored information with the information received, an approximation only similarity between the inputs to the prototype model. Thus, the possible existence of differences between the input and prototypes, models have become more flexible approach [7]. Prototype matching implementation to determine the scene categorization in this paper was applied to the Bayes' theorem [9], which interpretation of scene categorization base on grouping object on the some environment (discussion III).

Experiments conducted in this paper is intended to set the stage of searching scene categorization on SUN908 dataset is based on the probability distribution of objects found by the machine. The success of such understanding is highly dependent on how the system performs interpretation of a presence by integrating objects in the environment as a collection of objects (grouping) under the laws of gestalt. We examine the categorization of scenery to explore the relationship of objects with a name (label) which is in the scene on SUN908 Dataset.

III. ARCHITECTURE MODEL

The probability of the emergence of objects in one place can be used to determine the scene categorization. As an illustration, a semantically room we call the bedroom if there are fixtures in the room to sleep, as well as a living room, study room and others. Object relations is closely associated with the environment, environmental categorization can be identified simply by looking at the objects around the environment.

The scene categorization is annotated labels on a picture or a video that represents some place where people can move and navigate. In this paper, the entire sample category used is defined as a set ***S***. The set of objects scattered on the category defined as the set ***O***.

$$\boldsymbol{S} = \{\boldsymbol{s_i} = \boldsymbol{u},\ \boldsymbol{u} € \boldsymbol{V}(\boldsymbol{G})\} \qquad (1)$$

$$\boldsymbol{O} = \{\boldsymbol{o_i} = \boldsymbol{v}, \boldsymbol{v} € \boldsymbol{V}(\boldsymbol{G})\} \qquad (2)$$

S is a set that consists of the scene name that is represented node (vertex) of the knowledge graph. While ***O*** is the set consisting of the names of objects that are represented on the graph as a node that can have a relation on ***S***.

DATASET908 have a number of images that have been annotated (labeled) manually as well as images that have not been labeled (not annotated). So that the accuracy of the model was developed to minimize errors, the number of images extracted defined as a matrix ***A***, namely a matrix containing $\boldsymbol{a_i}$ as a matrix element contains the number of images that have been annotated version of the set ***S***. So $\boldsymbol{a_i}$ is a representation of ***wu(G)*** is the property ***u*** node $(\boldsymbol{s_i})$ that contains a value that represents the number of images anonotated

$$. \boldsymbol{A} = \{\boldsymbol{a_i} = \boldsymbol{wu}(\boldsymbol{G})\} \qquad (3)$$

The relation between ***O*** to ***S*** have different values, for example, a chair could be in the living room or dining room, but the number that often appears on each different space. So defined matrix ***B*** that represents the relation $\{\boldsymbol{o_i} \rightarrow \boldsymbol{s_i}\}$.

Thus, ***B*** is a matrix that contains the value (weighting) the relation of the objects $\boldsymbol{o_i}$ which has a relation to the scenes of $\boldsymbol{s_j}$.

$$\boldsymbol{B} = \{\boldsymbol{b_{ij}} = \boldsymbol{w} = \boldsymbol{u} \rightarrow \boldsymbol{v}\} \qquad (4)$$

As the object distribution matrix, the matrix B is strongly influenced by the number of images that have been annotated, the greater the value of the number of images that annotated then the chances will be changed, so as to equalize all knowledge of the normalization ***B*** $(\overline{\boldsymbol{B}})$ by multiplying ***B*** by $\{\frac{1}{A^T}\}$.

$$\overline{\boldsymbol{B}} = \{\overline{\boldsymbol{b_{ij}}} = \tfrac{1}{A^T}\boldsymbol{B}\} \qquad (5)$$

The model developed in this paper adopts the theory of probability that is a value used to measure the level of an occurrence that is random and often referred to the opportunities or possibilities. Probability is generally a chance that something will happen. Based on the understanding that it must be assumed that a set of objects have the opportunity to predict these objects represent a place or environment in which people can move and navigate.

The formulations used in this paper distribution probabilities represent random objects often appear in a scene using ***C*** matrix definition that contains the value of Probability $(\overline{\boldsymbol{b}}_{\boldsymbol{ij}}|\boldsymbol{a_i})$.

$$\boldsymbol{C} = \{\boldsymbol{c_{ij}} = \boldsymbol{P}(\overline{\boldsymbol{b}}_{\boldsymbol{ij}}|\boldsymbol{a_i})\} \qquad (6)$$

Object observation that appears on a drawing conducted by the observer, based on the label is entered into the system by the observer, then the definition of a group of objects is defined as ***Π*** is a matrix that identifies a set of managed objects discovered during observations of a scene (image).

$$\boldsymbol{\Pi} = \{\boldsymbol{\pi_i}\} \qquad (7)$$

Here in after ***Π*** multiplied by a ***C*** that were previously done to normalize the value by identifying the largest value of as objects that have the greatest quantity. This is done so that the legal opportunities ***0> = P <= 1.***

$$\boldsymbol{R} = \frac{\mathbf{1}}{\mathbf{max}_{\pi_i € \Pi}(\pi_i)}\boldsymbol{\Pi C} \qquad (8)$$

Inference being done to find the greatest value of matrix multiplication $\boldsymbol{r_i}$ observations using **λ** value.

$$\boldsymbol{\lambda} = \mathbf{argmax}_{r_i € R}(\boldsymbol{f}(\boldsymbol{r_n^{i=1}})) \qquad (9)$$

Furthermore, to get the label of a category based on value λ in scene of the craft following a simple algorithm:

Lable Transfer Algoritm

```
$R = \frac{1}{\max_{\pi_i € \Pi}(\pi_i)} \Pi C$
$\lambda$=0
Loop i=1 to R.length
    If $\lambda < r_i$
        $\lambda \leftarrow r_i$
        x=i
    End If
End Loop
Out.SceneCaategory is $S_x$
```

Simplification of the development of the model in this study is to define the scene category as a collection of objects (***Π***) as the Law of Perception, which has a reference value (principle) as the basis for the perception of the machine (***λ***).

Experiments were carried out in this paper to identify the level of accuracy of the system that uses the model developed arranged in the form diagram in Fig. 3

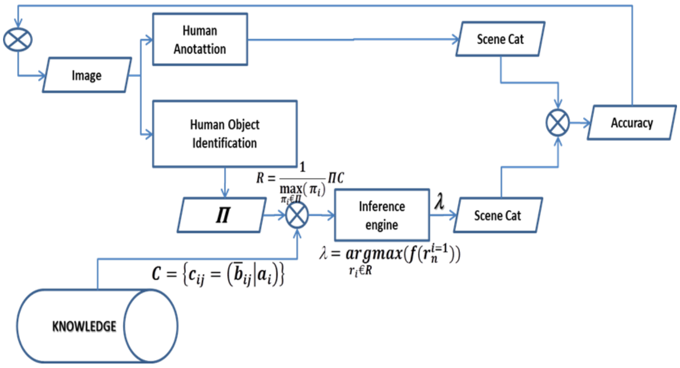
Human Anotattion
Scene Cat
Image
Human Object Identification
Accuracy
$R = \frac{1}{\max_{\pi_i € \Pi}(\pi_i)} \Pi C$
Π
Inference engine
λ
Scene Cat
$\lambda = argmax_{r_i € R}(f(r_n^{i=1}))$
$C = \{c_{ij} = (\overline{b}_{ij}|a_i)\}$
KNOWLEDGE

Fig 3. Diagram Model

IV. DISCUSSION AND EXPERIMENT

To test the knowledge representation and categorization scene using probability matrix, this paper uses dataset SUN908. The number of categories used are bathroom, bedroom, dining room, living room, playroom and television room (TABLE I) as many as 1045 images and representing it in Neo4j Graph Database Systems 2.3.2.

TABLE I
NUMBER OF COTEGORIES

No	Category	Number of Image
1	bathroom	280
2	bedroom	286
3	dining room	286
4	living room	154
5	playroom	37
6	television room	2

Further develop tools that an application for the implementation of the model that was developed involving 10 observer as Human Anotation with a distribution in 1045 the number of images as in TABLE II.

TABLE II
NUMBER IMAGE

Observer	bathroom	bedroom	dining room	living room	playroom	television room	Total Image
A	**26**	**26**	**26**	**13**	**7**	**2**	100
B	26	26	26	26	7	-	111
C	26	26	26	26	-	-	104
D	26	26	26	26	-	-	104
E	26	26	26	13	10	-	101
F	26	26	26	26	-	-	104
G	36	26	26	13	6	-	107
H	39	26	26	11	7	-	109
I	26	39	39	-	-	-	104
J	23	39	39	-	-	-	101

From TABLE II, not the whole picture is observed, only 1012 images annotated than 1045 images. There are 33 images are missing or not annotated with end result as in TABLE III.

TABLE III
NUMBER OF IMAGE ANNOTATED

bathroom		bedroom		dining room		living room		playroom		television room	
Human	Machine	Human	Machine	Human	Machine	Human	Machine	Human	Machine	Human	Machine
24	25	25	21	25	43	14	6	7	3	5	2
28	28	23	23	27	35	23	14	6	5	5	7
26	26	22	18	25	28	22	18	2	2	3	8
24	26	22	26	26	28	24	20	0	0	4	0
26	25	25	26	26	31	12	6	10	9	1	3
26	26	32	33	28	38	11	3	1	0	2	0
36	35	30	26	33	38	1	1	0	0	0	0
30	34	26	15	23	26	12	9	7	5	2	11
26	22	39	35	35	43	0	0	0	0	0	0
20	23	38	31	37	43	4	2	0	0	1	1

Because in this experiment using inference to determine the category in addition to a scene of recall and precision, also used the measurement accuracy. Recall is is the proportion of Real Positive cases that are correctly Predicted Positive and Precision denotes the proportion of Predicted Positive cases that are correctly Real Positives [10]. Accuracy is defined as the degree of closeness between the predicted value to the actual value.

From the experimental results gathered, then calculated precision, recall and accuracy of performance of systems that implement the developed model (TABLE IV and Fig 4).

TABLE IV
PERFOMANCE OF SYSTEM

TEST	Bathroom	Bedroom	Dining room	Living room	Playroom	Television room	Average
Precision	0,96	0,98	0,82	1,00	1,00	0,60	0,89
Recall	0,98	0,88	1,00	0,63	0,78	0,90	0,86
Acuracy	0,98	0,96	0,92	0,95	0,97	0,97	0,96

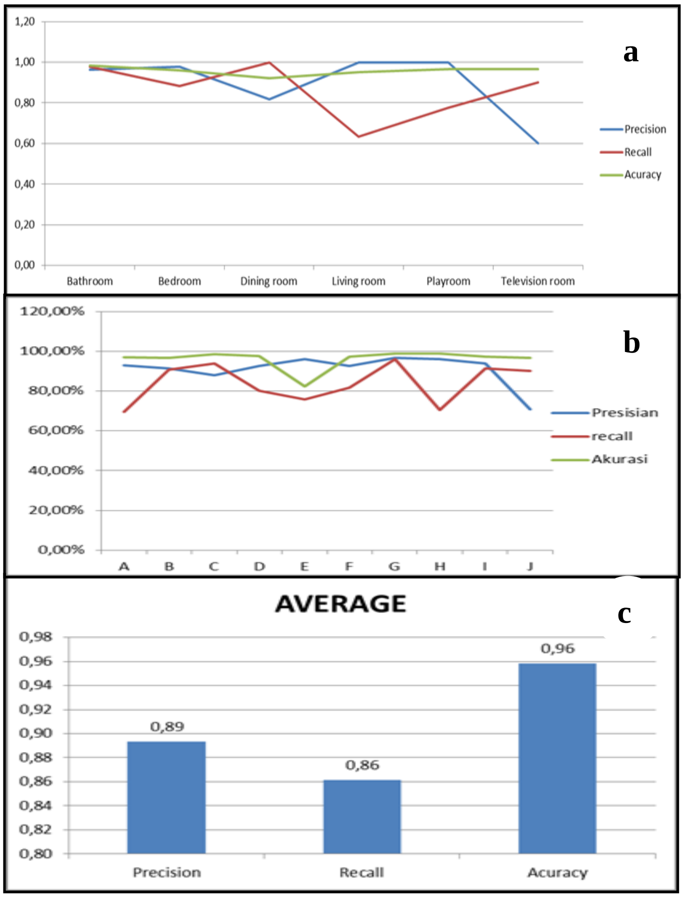

Fig 4. Performance of system a) measurement (precission, recall and acuracy) base on scene categories b) measurement base on observer c) averages of measurement.

On the Fig 4.a, recall value is the lowest in the category of living room, while the precision found in the television room, the association of these two categories is ambiguity on the difference in living room and television room, as well as the lack of a database on the category of television rom which only has 2 pieces picture (TABLE I)

Recently. From the whole set of experiments there are some mistakes annotations by humans, but can be repaired by the system (Figure 5). This is caused by the machine's ability that appear to group objects in the image into a unity or feature that has reference values used to identify categories of scenes (equation 7).

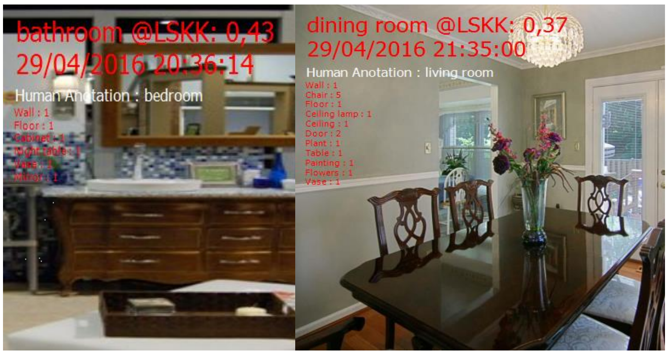

Figure 5. Errors human identification, but correct identification of the machine

V. CONCLUSION

Conclusion of this experiment is a collection of objects very big influence on categorization, in other words a space can change its function if objects are placed (found) is different. Of the six categories of scene is used as a knowledge base, television room has a minimal dataset (2 pictures) so that recall is low and can affect the accuracy of the prediction.

With the models and algorithms developed indicate a high degree of prediction accuracy on static images (single image) and use the observer as a user to provide input to the system of objects that appear.

Future Work: In a subsequent study, we will tests on streaming video and uses object detection algorithm, so expect the machine can perform independently and automatic categorization

ACKNOWLEDGMENT

We would like thanks and appreciation to the anonymous for providing comments and suggestion.

REFERENCES

[1] C. Henson, “A Semantics-based Approach to Machine Perception,” Computer Science Wright State University, USA, 2013.

[2] P. Barnaghi, F. Ganz, C. Henson, dan A. Sheth, “Computing Perception from Sensor Data,” in *IEEE Sensors Conference*, 2012.

[3] G. Patterson, C. Xu, H. Su, dan J. Hays, “The SUN attribute database: Beyond categories for deeper scene understanding,” *Int. J. Comput. Vis.*, vol. 108, hal. 59–81, 2014.

[4] J. Xiao, J. Hays, K. a. Ehinger, A. Oliva, dan A. Torralba, “SUN database: Large-scale scene recognition from abbey to zoo,” *2010 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, hal. 3485–3492, Jun 2010.

[5] J. Xiao, B. C. Russell, J. Hays, K. a Ehinger, A. Oliva, dan A. Torralba, “Basic level scene understanding: from labels to structure and beyond,” *SIGGRAPH Asia 2012 Tech. Briefs*, hal. 36:1--36:4, 2012.

[6] J. Xiao, K. a. Ehinger, J. Hays, A. Torralba, dan A. Oliva, “SUN Database: Exploring a Large Collection of Scene Categories,” *Int. J. Comput. Vis.*, 2014.

[7] K. M. Galotti, “Cognitive Psyghology in and Out of the Laboratory,” in *Cognitive Psychology In and Out of the Laboratory Electronic Version*, FIFTH EDIT., Carleton College, Minnesota.: SAGE Publications, Inc, 2013, hal. 39–64.

[8] R. Brunelli, *Template Matching Techniques in Computer Vision: Theory and Practice*, FIRST. John Wiley & Sons, Ltd, 2009.

[9] E. N. Hájek, Alan, Zalta, “Interpretations of Probability,” *Stanford Encyclopedia of Philosophy*. 2012.

[10] D. M. W. Powers, “Evaluation : From Precision , Recall and F-Factor to ROC , Informedness , Markedness & Correlation,” *Int. J. Mach. Learn. Technol.*, vol. 2, no. December, hal. 37–63, 2011.